The Ethics of Automated Vehicles

Geoff Keeling

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Arts.

April 2020

Word Count: 69,872

Abstract

My thesis is about the morality of automated vehicle (AV) decisions. What is the relevance of the trolley problem to AV decisions? What is the morally right method for AVs to distribute harms or risks of harm between the parties in collisions? Who is morally responsible for harm caused in collisions? How does the risk of harm to road-users trade-off against the AV's prudential goal of getting to its destination in reasonable time? What is the morally right amount of caution for AVs to exercise when uncertain about the classification of proximate objects? I will answer these questions by developing a deontological account of permissible killing for AVs and a deontological account of permissible risk-imposition. I also argue against some of the rival answers to these questions; and in doing so, attempt to show that my view does a better job than its rivals at capturing our considered moral judgements.

The view I defend holds that the AV is morally permitted to kill or harm a road-user in an unavoidable collision if, and only if, and because, its passenger is morally permitted to kill or harm that road-user in self-defence. In normal driving, the AV is morally required to moderate its speed so that it can safely negotiate modally close but improbable *what if* cases such as children running out into the road. Here is the plan. Chapter 1 is a literature review. Chapter 2 defends my use of trolley cases. Chapters 3 and 4 argue against two rival non-consequentialist views, Filippo Santoni de Sio's (2017) legal-philosophical view and Derek Leben's (2017) Rawlsian view. Chapter 5 presents my deontological theory. Chapter 6 presents my account of risk-imposition, and develops an illustrative formal decision-procedure for a simple mundane road-traffic scenario modelled as a Markov Decision Process.

Acknowledgements

I owe an enormous debt of gratitude to my supervisors, Richard Pettigrew and Brad Hooker, for their kindness and encouragement at every stage of my academic development. I have learned so much from these two. I cannot thank them enough. I am also deeply grateful to Michael Hauskeller for his supervision and mentoring in the early stages of the PhD; and to Rune Nyrup, who has supervised me since June 2019 at the Leverhulme Centre for the Future of Intelligence in Cambridge. I am immensely fortunate to have had such wonderful mentors throughout my PhD.

I am lucky to have some amazing friends. Farbod Akhlaghi-Ghaffarokh has pushed my dialectical frontier further than I could have imagined; and has been a constant source of inspiration and friendship. Nick Axten reminds me each time we talk that philosophy begins in wonder. In addition, I owe so much to Yousuf Bhyat, Chris Burr, Pavan Chaggar, Leia Hopf, Daniel Jones, Isaac Kean, Tim Keeling, Aadil Kurji, Arsham Nejad Kourki, Niall Paterson, and Shaun Stanley. Bristol's graduate community has also been brilliant. In particular, thanks to Lize Alberts, Alejandra Casas Munoz, Nemo D'Qrill, Elle Garner, Lucy James, Jack Lane, Jed Martin, Nick Ormrod, Gareth Pearce, João Pinheiro, Nicos Stylianou, Demyan Vakhrameev, Jacqui Wallis, James Wilson, Aiden Woodcock, and Kieran Woods.

I have some wonderful friends and colleagues in the ethics of automated vehicles. In particular, I must thank Katie Evans, Nick Evans, Filippo Santoni de Sio, Noah Goodall, Jeff Gurney, Johannes Himmelreich, Jamie Hodsdon, Ryan Jenkins, Derek Leben, Sven Nyholm, Pamela Robinson, Sarah Thornton, Carole Voulgaris, and Damian Williams. I have learned so much from discussions with these people and from their comments on my work. I owe special thanks to Noah, Nick, and Pamela for inviting me to present at the Automated Vehicles Symposium in San Francisco in 2018 and in Orlando in 2019. I have also benefited from many fruitful conversations with Kevin Baum, Georg Borges, Matt Clark, Juan Durán, Thomas Grote, Lauren Holt, Abhishek Mishra, Will McNeill, Brent Mittelstadt, Farzad Nozarian, Duncan Purves, Eva Schmidt, Sandra Wachter and Fiona Woollard.

The Faculty in the Department of Philosophy at Bristol have been amazing. I owe special thanks to Chris Bertram and Jason Konek. But I have learned so much

from conversations, questions, and comments from Catrin Campbell-Moore, Tzu Chien Tho, Ana Cretu, Josh Habgood-Coote, Max Jones, James Ladyman, Karim Thébault, Martin Sticker, Samir Okasha, and Alan Wilson. I am also eternally grateful to Tamar Hodos from the Archaeology and Anthropology Department for the interest that she has taken in my career and for so much helpful advice. Thanks must also go to the South, West, and Wales Doctoral Training Partnership for the Arts and Humanities Research Council Studentship that funded this venture.

Last, I thank my family. In his book *Ideal Code, Real World*, Brad says that 'my parents openly acknowledged that moral projects are often trying. Raising a son fascinated by a subject at once so difficult and emotionally charged as moral philosophy served as a case in point' (Hooker 2000: viii). For their sins, my parents have raised two sons that are fascinated by moral philosophy. This thesis is for my mother, Dr Julie Keeling, the person of whom I am most proud; and no doubt the person who is most proud of me. Thank you for everything. You are brilliant.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Publications

In line with Annex 5 of the Regulations and Code of Practice for Research Degree Programmes, some material in this thesis is reprinted in the following publications:

Keeling, G. Legal Necessity, Pareto Efficiency & Justified Killing in Autonomous Vehicle Collisions. *Ethical Theory and Moral Practice* **21,** 413–427 (2018).

Keeling G. Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles. In: Vincent Müller (ed). Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. SAPERE, vol 44. Springer, Cham (2018).

Keeling, G. Why Trolley Problems Matter for the Ethics of Automated Vehicles. *Science and Engineering Ethics* **26,** 293–307 (2020).

Keeling, G. Automated Vehicles and the Ethics of Classification. In: Ryan Jenkins, Tomas Hribek and David Cerny (eds.). *Autonomous Vehicle Ethics: Beyond the Trolley Problem*. Oxford University Press (Accepted).

Chapters 2, 3, 4, and 6 are 'publication chapters', in that substantial content from these chapters features in the above publications. Footnotes at the start of these chapters detail which publications the chapters correspond to. Chapters 1 and 5 are 'conventional chapters', in the sense that these chapters are unpublished.

Contents

1. The Ethics of Automated Vehicles

My thesis is about the morality of automated vehicle (AV) decisions. What is the relevance of the trolley problem to AV decisions? What is the morally right method for AVs to distribute harms or risks of harm between the parties in collisions? Who is morally responsible for harm caused in collisions? How does the risk of harm to road-users trade-off against the AV's prudential goal of getting to its destination in reasonable time? What is the morally right amount of caution for AVs to exercise when uncertain about the classification of proximate objects? I will answer these questions by developing a deontological account of permissible killing for AVs and a deontological account of permissible risk-imposition. I also argue against some of the rival answers to these questions; and in doing so, attempt to show that my view does a better job than its rivals at capturing our considered moral judgements.

The point of this chapter is to set the scene. I will survey the landscape, making precise the main contours in the dispute over the morality of AV decisions. My aim is not to provide an impartial treatment of the literature. What follows is better seen as an opinionated guide. In §1.1, I cover some preliminaries. In §1.2 and §1.3, I discuss the *moral design problem*. This problem asks what morality requires in dilemmatic collisions with respect to balancing harms or risks of harm between the affected parties. In §1.4, I discuss the *blame problem*, i.e. who is responsible for harm or property damage caused by AVs. In §1.5, I discuss the *risk-imposition problem*, which concerns the degree of risk that AVs are morally permitted to impose on other road-users. Last, in §1.6, I provide a roadmap for the thesis. One discussion that I shall not comment on in detail is the relevance of the trolley problem to the ethics of AVs. I shall provide a complete treatment of this issue in Chapter 2.

1.1. Preliminaries

I shall make three points. First, vehicle automation is a matter of degree. What this means is that a vehicle can be *more* or *less* automated. The Society for Automotive Engineers (2014) has a framework for classifying different levels of automation.[1] I shall use the term 'AV' to mean Level 5 automated vehicles in accordance with this framework. These vehicles are fully automated. They can drive in all the same roadway and environmental conditions as a competent human driver. There is no need for human supervision or intervention. These vehicles need not have manual controls. To keep things simple, I shall assume that AVs have no manual controls.

Second, AVs will be on our roads at some point in the future. I am unsure when. The timeline is hard to predict because the widespread adoption of AVs depends on certain technical developments, consumer demand, and the regulatory approach of different states. Todd Litman (2020: 14) predicts that Level 5 AVs will be available with a large price premium by the 2030s, but that market saturation will not occur until the 2070s. I do not think that it matters for our purposes *when* AVs are expected to saturate the vehicle market. What is important is that at some point in the short- or medium-term AVs will become commonplace on public highways.

Third, AVs are widely expected to reduce the number of road-traffic deaths. Janet Fleetwood (2017: 532) notes that around 90% of the 29,000 traffic fatalities that occur each year in the United States are the result of human errors such as being drunk or asleep at the wheel (c.f. NHTSA 2016: 26). Because AVs do not make errors like these, Fleetwood argues that the widespread adoption of AVs could reduce the number of road-traffic deaths significantly. I suspect Fleetwood is correct. But caution is required. As Noah Goodall (2017: 496) puts it, 'Just as airbags save many, they also kill a few that would otherwise not have died.' What is at issue here is that we do not know what sorts of fatal errors AVs will make, or how frequently these errors will be made. Indeed, Bryant Walker Smith (2012) has used human crash statistics to argue that an AV would need to drive without a fatal

[1] Here is a brief overview of their taxonomy: Level 0 vehicles are such that a human driver performs all functions. Level 1 vehicles have at most one driver-assist feature. Level 2 vehicles are 'partially automated', in the sense that they have multiple driver-assist features that can run concurrently. Level 3 vehicles have 'conditional automation', in that there is an automated driving mode that can be used under certain conditions. But a human driver is required to supervise and respond to a request to intervene. Level 4 vehicles have 'high automation', i.e. there is an automated driving mode that can be used under certain conditions; and human supervision or intervention is not required during automated driving mode. Finally, Level 5 vehicles have full automation, in the sense that these vehicles have an automated driving mode that operates at all times on the road which can manage any roadway or environmental conditions that a human driver can manage (SAE 2014).

crash for 300 million miles before we can be 99% confident that the AV has fatal collisions less frequently than normal vehicles (c.f. Goodall 2014: 59). Hence precise claims about the safety benefits of AVs should be taken with a pinch of salt.

1.2. The Moral Design Problem

I now turn to the first ethical problem. What does morality require in unavoidable collisions in which the AV cannot avoid imposing a risk of harm on at least one person; and a choice is required about how to allocate harms or risks of harm between multiple persons whose interests are in conflict? What is the morally right way for the AV to balance the competing interests of the different parties? Which moral principles best explain the moral status of the AV's acts in these collisions? I call this the *moral design problem* (Keeling 2017, 2018a, 2018b, 2020).

Why care about this problem? The standard way to motivate the moral design problem is to provide examples of dilemmatic collisions that AVs could face. Patrick Lin (2016: 76) gives an example in which an AV is driving on a narrow road that has a steep cliff-edge on one side. A schoolbus comes around the bend on the wrong side of the road. The AV can either brake, in which case it will collide with the bus and impose a serious risk of harm on its passenger and on the occupants of the bus. Or it can swerve-off the cliff, avoiding the collision, but imposing a serious risk of death on its passenger (c.f. Goodall 2014: 60-61). The next move is to point out that in cases like these there are moral considerations that bear on what the AV ought to do. For example, it might make a moral difference that the AV's passenger is guaranteed at least a serious risk of harm on both options, but that the occupants of the schoolbus will be harmed only if the AV brakes. Likewise, it might make a difference that the schoolbus is driving on the wrong side of the road, and thus its driver is imposing an undue risk of harm on other road-users. The final move is to argue that because it is unclear what the AV morally ought to do in cases like these, a solution to the moral design problem is required before AVs are made available.

I think that caution is required here. This argument suggests that the reason to care about the moral design problem is that we need to tell a plausible story about how AVs ought to behave should a dilemmatic collision arise. This point invites several criticisms. First, it stakes the motivation for the moral design problem on an empirical claim about whether or not AVs will encounter dilemmatic collisions; and some people think that these collisions are impossible or unlikely to arise in practice (Himmelreich 2018; Goodall 2016, 2019; Keeling et al. 2019). Second, even

granting that dilemmatic collisions are possible, it is unclear that we have good reason to focus on what morality requires in these presumably rather unlikely cases. Perhaps what matters more are the moral dilemmas that arise in normal driving, such as those that concern permissible risk-imposition (Himmelreich 2018). Thus it is not ideal to rest the case for the moral design problem on the claim that it is in principle possible that AVs might encounter dilemmatic collisions in practice.

Can we do better? I shall discuss this issue in Chapter 2. But for the time being, I shall motivate the moral design problem as follows. When Frances Kamm tries to motivate the use of trolley problems in the ethics of killing, she claims:

> On the basis of what I have said so far, we can see why the cases that have been the focus of attention in trolley problem discussions seem artificial and unrealistic. They are specifically constructed, like scientific experiments, to distinguish among and test theories and principles (e.g., consequentialist vs. nonconsequentialist theories) because one theory or principle would imply the permissibility of conduct that the other theory or principle would deny. Using our intuitive judgments about which implications for cases are correct helps us decide among, and also revise, theories and principles (Kamm 2016: 13).

Kamm's point is that artificial cases help us to decide between moral theories, and to test which features of acts make a difference to the moral status of those acts. Ideal cases allow us to hold fixed all the features of the case bar one and then test whether permuting that feature makes a difference to our moral judgements. The aim of these thought experiments is not to work out what we ought to do in the unlikely event that we are a train driver forced to decide whether to hit five people on the track ahead, or divert the train so as to kill one person on the other track. The same point holds for thought experiments in the ethics of AVs. I suspect that dilemmatic collisions could arise. I do not know how frequently. But it seems to me that the principal reason to care about the moral design problem is that it helps us to determine the morally relevant facts in noisy real-world cases. I know that some people disagree with my opinion. I shall respond to these people in Chapter 2.

I shall now consider some answers to the moral design problem. As one might expect, the usual suspects have emerged: the self-interest theory, act utilitarianism, contractarianism, contractualism, and deontology. Before I consider what these five theories say about the morality of AV collisions, I should flag an important concern: Do AV collisions present a new moral challenge? Or do AV collisions instantiate the same old considerations from the ethics of killing, harming, and imposing risks? This question has an important role to play later on. I believe that AVs do not

present a *sui generis* moral context. These really are the same old problems. That does not mean that existing moral theory can be applied straightforwardly to this novel context. There is much work to be done in discerning the morally relevant considerations whenever an existing theory is applied to a new domain. Others disagree. These people believe that AVs present fundamentally different moral challenges; and some of these people have defended novel moral theories for AVs. I shall discuss these other views once I have talked about the usual suspects. I will provide a full treatment of this issue in Chapter 5 when I develop my own theory.

1.2.1. Self-Interest Theory

According to what I shall call the *self-interest theory*, AVs ought to maximise the (expected) welfare of their passengers in collisions (Keeling et al. 2019: 54–56; see also Contissa et al. 2017: 370).[2] The expected welfare for a given act is the sum of the welfare that the passengers receive if the act is performed under different possible states of the world multiplied by the probability of each state. The self-interest theory holds that the AV ought to choose that act in a collision which will result in at least as much (expected) welfare for its passengers as all the alternatives.

It is important to distinguish the self-interest theory as a decision procedure and as a criterion of rightness (c.f. Brink 1986: 421). When self-interest theory is taken to be a criterion of rightness it holds that what explains the rightness of the AV's acts is facts about the (expected) welfare of its passengers. The AV's act is right if, and only if, and because, it produces at least as much (expected) welfare for its passengers as each of the available alternatives. The self-interest theory as a decision-procedure holds that in collisions the AV morally ought to deliberate like this: evaluate the (expected) welfare to the passengers on each alternative, and choose the alternative that maximises their (expected) welfare. The self-interest theory faces some interesting puzzles. I shall outline some of these puzzles here.

First, the self-interest theory as a criterion of rightness does not recommend the self-interest theory as a decision-procedure. The problem is that, in some cases, the AV's acting in the passenger's best interests is not in the passenger's best interests

[2] Note that self-interest theory as a solution to the moral design problem differs slightly from self-interest theory in normative ethics. On this view, each of us has the substantive moral aim of making our lives go as well as possible (Parfit 1987: 3). Self-interest theory as I am formulating it as an answer to the moral design problem holds that the AV has the substantive moral aim to make things go best for its passengers in collision scenarios.

(c.f. Gogoll and Müller 2017; see also Parfit 1987: 56-62). This is true in collision scenarios involving two AVs that instantiate the prisoner's dilemma. Consider,

	B Continues	**B Swerves**
A Continues	A Injured B Injured	A Dies B Unharmed
A Swerves	A Unharmed B Dies	A Badly Injured B Badly Injured

Here there are two AVs. One contains person A. The other contains person B. The AVs can either swerve or continue. The Nash Equilibrium is for both AVs to swerve, as neither A nor B benefits from changing strategy in this arrangement. But if the AVs both swerve, A and B will be badly injured. Thus if the AVs act in the best interests of their passengers, A and B will be worse-off than if the AVs had coordinated to produce the optimific outcome on which both A and B are injured. Thus the self-interest theory is less transparent than we might imagine.[3] It might be true that my AV ought to prioritise my safety. But sometimes this is best done by cooperating with other AVs to ensure the mutual safety of all the affected parties.

Second, the self-interest theory requires a plausible story about why it is that the rightness of the AV's acts depends only on facts about passenger welfare. What seems like the most plausible view is that AV companies have an agent-relative obligation to prioritise the safety of their passengers (Keeling et al. 2019: 54-56). Thus the maximisation of passenger welfare is what matters because AV companies stand in certain historical relations to the passengers which ground an obligation to prioritise the safety of those passengers. This is not yet a plausible story. What is not clear is *which* historical facts ground the agent-relative obligation, and *why* this obligation outweighs the general duty not to cause harm to other road-users.

This explanatory burden is challenging. Part of the problem is that the agent-relative obligation to assign lexical priority to passenger welfare over the welfare of other road-users leads to absurd consequences in cases involving risk. Suppose that the AV can avoid killing a pedestrian only if it imposes a 1% risk of death on its passenger. Here the self-interest theory holds that the AV ought to kill the pedestrian so as not to impose a 1% risk of death on its passenger. This is clearly

[3] Here I am assuming that each AV has no means to predict what the other AV will do. If AVs could make reasonable probabilistic estimates about what other AVs will do in crashes, then it may not be rational for the AVs to opt for the Nash Equilibrium.

false. I do not see which feature of the historical relationship between AV companies and their passengers is sufficient to ground extreme partiality of this sort.

What strikes me as the best option is for proponents of the self-interest theory to moderate their view, and hold that the AV ought to assign *greater*, but not *complete*, priority to the welfare of its passengers (Keeling et al. 2019: 56-58). This might be called the *passenger partiality view*. This view is more plausible than the self-interest theory at least insofar as it avoids the obscene counterexamples involving risk. But this view must in turn provide an account of how much partiality the AV is permitted to exercise towards its passengers; and which features of the company/passenger relationship render this amount of partiality permissible.

The amount of partiality that an AV is morally permitted to exercise towards its passengers is, I think, quite limited. Suppose that I hire a bodyguard to protect me. Presumably, there are limits to what the bodyguard is morally permitted to do to other people in my defence. These limits are determined by the rules for the use of other-defensive force. The standard view is that the bodyguard's other-defensive permissions are set by the amount of force that I am in principle permitted to use in self-defence (Frowe 2014: 98f). Roughly, I am permitted to use proportionate force in response to an unjust threat. Thus whilst the bodyguard has an obligation to protect me grounded in their contractual promise to do so, there are independent limits to the amount of force that the bodyguard is morally permitted to exercise in fulfilment of this promise. The passenger partiality view has a lot of work on its hands if it holds that certain features of the AV company/passenger relationship make it the case that the AV is morally permitted to *exceed* the normal other-defensive moral permissions. However, if the claim is that AVs can be partial to passengers within the bounds of other-defensive permissions, then the view is not all that interesting.[4] I shall explore this problem in more detail in Chapter 5.

[4] It is difficult to think of products other than AVs for which this would be an issue. This is because very few products exist where those products have sufficient autonomy to exercise partiality towards their owners. The closest example I can think of is *booby traps* installed in the home to be used in defence against burglars. The relevant statutory law is the 2013 amendment of s.76 of the Criminal Justice and Immigration Act 2008. This amendment relaxed the requirement on homeowners to use reasonable force against intruders, and held that homeowners would be given the benefit of the doubt for unreasonable force used. But any force used that counts as *grossly disproportionate* is not covered by self-defensive legal permissions. See *R v Martin* [2001] 1 Cr App R 27 for guidelines on proportionate force in home-invasion cases. Based on the existing law, I imagine that any booby trap device used must not be grossly disproportionate in line with self-defensive permissions.

1.2.2. Utilitarianism

Jeff Gurney (2016: 211-17) and Jean-François Bonnefon et al. (2016: 1573) have floated, but not endorsed, an act utilitarian solution to the moral design problem (c.f. Gerdes and Thornton 2015: 91; Gogoll and Müller 2017; Keeling 2018a: 424). Like the self-interest theory, act utilitarianism can be formulated as a decision-procedure and as a criterion of rightness. The latter holds that an act is right if, and only if, and because, it produces at least as much (expected) welfare impartially considered as all the alternatives. The former holds that in collisions AVs ought to deliberate as follows: evaluate each act in terms of (expected) aggregate welfare, and choose the act which maximises (expected) welfare impartially considered.

One puzzle for the act utilitarian view is that if maximising welfare impartially considered is what matters morally, then there is no principled reason to restrict the scope of the ethical analysis to AV collisions. Presumably, the correct level of utilitarian analysis is the overall welfare consequences of AVs in general having different crash algorithms. Perhaps consumer demand for utilitarian AVs is non-existent. That is to say that people do not want their AVs to sacrifice them and their families in order to save the greater number in collisions. If it is also true that AVs are much safer than normal cars, then mandating utilitarian crash algorithms is likely to result in more road-traffic fatalities than mandating some other non-utilitarian crash algorithm (Nyholm 2018a: 6; see also Bonnefon et al. 2016). This is because mandating utilitarian crash algorithms under these assumptions would most likely prevent the adoption of AVs over manually driven vehicles. Thus an act utilitarian decision-procedure is appropriate on utilitarian grounds only if the general implementation of that algorithm will maximise aggregate welfare.

This puzzle is familiar from Derek Parfit's discussion of consequentialism. In the following passage, Parfit uses 'C' to denote consequentialism. Consider,

> C gives us one substantive moral aim: that history go as well as possible. Does it also give us a second substantive aim: that we never act wrongly? On the best known form of C, Utilitarianism, the answer is No. For Utilitarians, avoiding wrong-doing is a mere means to the achievement of the one substantive moral aim. It is not itself a substantive moral aim (Parfit 1987: 37).

The problem is now less puzzling. The mistake is that act utilitarianism does not assign to AVs the substantive moral aim of never acting wrongly by the lights of utilitarianism. That is to say that act utilitarianism does not require AVs to act so as to maximise aggregate welfare in collisions. What act utilitarianism requires

is that we act so as to produce the most (expected) welfare in the long-run. If this means implementing non-utilitarian AV collision algorithms, then so be it.

What follows from all this? First, that the utilitarian criterion of rightness need not recommend a utilitarian crash algorithm for AVs suggests that utilitarianism as a solution to the moral design problem is widely misunderstood. Consider,

> The basic approach of optimal control – choosing the set of inputs that will optimize a cost function – is directly analogous to consequentialist approaches in philosophy. If the ethical implications of an action can be captured in a cost function, as preference utilitarianism attempts to do, the control inputs that optimize that function produce the ideal outcome in an ethical sense. Since the vehicle can re-evaluate its control inputs, or acts, to produce the best possible result for any given scenario, the optimal controller operates according to the principles of act consequentialism (Gerdes and Thornton 2015: 91).

This is not right. Here Sarah Thornton and J. Christian Gerdes offer an act utilitarian decision-procedure. But this algorithm is not 'directly analogous to consequentialist approaches in philosophy.' The act utilitarian is unlikely to insist that AVs behave in accordance with an act utilitarian decision-procedure. What is important to them is not that AVs never act wrongly, but that we adopt an ethics policy for AVs in general that maximises welfare impartially considered.

It might be objected that what I am presenting is rule utilitarianism and not act utilitarianism. According to rule utilitarianism, an act is right if, and only if, and because, it follows from a set of rules the general internalisation of which by all or most people would be expectably best in terms of aggregate welfare relative to any other set of rules (Hooker 2000: 32; c.f. Brandt 1959, 1963; Harsanyi 1982, 1993). This objection misunderstands the distinction between act and rule utilitarianism. I am not claiming that act utilitarianism implies that AVs act rightly in collisions just in case they act in accordance with a set of rules the general implementation of which across all AVs would be optimific. That would be smuggling in a form of rule utilitarianism. What I am claiming is that act utilitarians do not care if AVs act wrongly by their standards in particular collisions, because what these people take as the substantive moral aim is the maximisation of welfare impartially considered. Hence act utilitarians favour the crash algorithm or set of crash algorithms that if implemented in AVs would maximise aggregate welfare all things considered.

So which algorithm would act utilitarians endorse? This is a difficult question. Shelly Kagan (1992: 226) tells us that it is rare for foundational moral theories like utilitarianism to provide 'a tradeoff schedule in complex cases involving conflicting

factors.' He continues that 'in practice foundational [moral] theories are virtually never worked out in this kind of detail.' Kagan's point here is that it is not possible to read-off a solution to applied ethical problems like the moral design problem from the foundational tenants of theories like utilitarianism. Really, I think that asking what utilitarianism implies about the moral design problem gets things the wrong way around. Once we have a systematic account of our considered moral judgements about how AVs ought to behave in collisions, the utilitarian will ask how those judgements can be explained by a utilitarian criterion of rightness.[5]

I appreciate that this answer is not very helpful. We have a practical ethical problem to solve. The question is: Can utilitarianism help us solve that problem? One thing I can say in answer to this question is that I think that an act utilitarian crash algorithm fails to capture our considered judgements about what is morally required in AV collisions. This is because such an algorithm does not take into account factors that we take to be morally relevant. For example, imagine that five hooligans decide to joy ride along a narrow mountain road. Jane is travelling in her AV along the same road. The hooligans drive around a bend on the wrong side of the road. Jane's AV can either swerve-off the side of the mountain, killing Jane and leaving the hooligans unharmed; or it can apply the brakes, in which case the five hooligans will die and Jane will be badly injured. The optimific outcome is the one that obtains if the AV kills Jane. But Jane has a moral complaint. The hooligans are responsible for the collision. Hence it is unfair that she should pay the extra cost.

Obviously, defenders of the act utilitarian decision-procedure can dig their heels in and insist that who is responsible for the collision is morally irrelevant. But this argument is difficult to sustain in the absence of a utilitarian criterion of rightness. That is to say that unless what ultimately explains the rightness or wrongness of the AV's acts in collisions is facts about the maximisation of aggregate welfare, it is hard to sustain the view that facts about moral responsibility, what is fair, and so on, should not be taken into account in the AV's deliberations. The trouble is that the act utilitarian criterion of rightness renders it an open question whether the act utilitarian decision-procedure is correct. In fact I suspect that few people would be

[5] Of course, utilitarians of different stripes will disagree. What I have in mind is the flavour of utilitarianism standardly espoused in normative ethics by people like Brad Hooker (2000) and Roger Crisp (2006). In contrast, utilitarians like Peter Singer and R.M. Hare have addressed some applied ethical questions through the application of utilitarian (or 'Kantian utilitarian') principles to specific contexts (e.g. Hare 2002: esp. 221-223). So, it is not out of the question for utilitarians to derive a solution to the moral design problem from a basic utilitarian principle. But the salient point is that applied ethical theorising of this sort does not consist in the naïve specification of a moral betterness ordering that ranks uncertain prospects from best to worst in terms of expected aggregate welfare. There is a great deal of work to do to discern what foundational moral theories imply about applied ethical cases.

inclined to purchase utilitarian AVs, such that the general adoption of a utilitarian decision-procedure for AVs would not maximise welfare all things considered, as many people would die in preventable collisions arising from human error. Hence the strongest argument for the utilitarian decision-procedure lends greater support to non-utilitarian decision-procedures. This presents a serious challenge to the idea that AVs ought to be programmed with act utilitarian crash algorithms.

1.2.3. Contractarianism

According to contractarianism, the content of morality or the reason to be moral is explained by facts about how rational self-interested people would agree to regulate their behaviour to ensure mutual benefit. Thomas Hobbes (1651/2008) defended a form of contractarianism; and more recent proponents of the view include David Gauthier (1986) and Jan Narveson (1988). I now discuss a contractarian solution to the moral design problem developed by Jan Gogoll and Julian Müller (2017).

According to Gogoll and Müller, AVs ought to adopt an average utilitarian decision-procedure on which the AVs involved in the collision coordinate so as to choose the acts that maximise average utility across all the affected parties.[6] This decision-procedure is defended by appeal to a contractarian criterion of rightness. Gogoll and Müller argue that prior to being involved in a collision, it is rational for each person to want AVs to maximise the average welfare across all parties in any collision that they might be involved in. Though an average utilitarian solution to the moral design problem might leave some people worse-off *ex post*, i.e. after a collision has occurred; each person is best-off *ex ante* if they consent to the average utilitarian collision algorithm, i.e. before they are involved in a collision.

Gogoll and Müller make no reference to John Harsanyi's (1977) defence of utilitarianism in welfare economics (c.f. Hübner and White 2018). But the argument

[6] Obviously, the prescriptions of average and total utilitarianism are necessarily coextensive in same-number cases. But we can draw a hyperintensional distinction between the requirements of average and total utilitarianism nevertheless. It matters for Gogoll and Müller that mean utility is the object of maximisation because maximising average welfare across all parties is equivalent to maximising expected utility from an impartial standpoint. Provided they can motivate the moral relevance of the point of view on which the moral status of the AV's acts is invariant under transformations of the identities in the affected parties, average utilitarianism follows from basic assumptions about rationality. In contrast, total utilitarianism requires us to motivate the moral relevance of the totality of welfare. This is standardly done by appeal to impersonal moral good that is distinct from each person's prudential good. This is why total utilitarianism is susceptible to John Rawls' (1971: 26-27) separateness of persons objection, whereas average utilitarianism is not.

is more or less the same. Harsanyi's view is that the just distribution of goods in a society is that which a rational and impartial individual would choose. Formally, let each individual $i = 1,2,3, \ldots, n$ have a utility function u_i that provides a cardinal measure of their welfare, such that $u_i(d_k)$ is the utility of distribution $d_k \in D$ for the ith individual. The impartial point of view is one on which the moral value of a distribution is invariant under permutations in the identities of the individuals. Harsanyi imagines the impartial individual as being behind a *veil of ignorance*, such that they do not know which utility function u_i belongs to them. Harsanyi argues that a rational individual in these circumstances would assign probability $1/n$ to the event that each u_i is theirs, and then choose the distribution d_k that satisfies:

$$\underset{d_k \in D}{\operatorname{argmax}} \left(\frac{1}{n} \sum_{i=1}^{n} u_i(d_k) \right)$$

This formalism means: 'Go through each distribution d_k in the set D, and choose that which maximises average utility over all the citizens.' That is to say that for Harsanyi the morally best distribution is that which maximises average utility.

Despite the apparent similarity between Gogoll and Müller's solution to the moral design problem and Harsanyi's utilitarianism, there is a crucial difference between these theories that brings out a problem for Gogoll and Müller's view. It might be true that in the circumstances of Harsanyi's veil of ignorance it is rational to prefer the distribution that maximises average welfare. But this point lacks moral significance in the absence of a principled reason to think that what is rational to choose behind Harsanyi's veil of ignorance is indicative of the just distribution of goods. Harsanyi provides an interesting answer to this problem. Consider,

> Suppose somebody tells us: "I much prefer our capitalist system over any socialist system because under our capitalist system I happen to be a millionaire and have a very satisfying life, whereas under a socialist system I would be in all probability at best a badly paid minor government official." This may be a very reasonable judgment of personal preference from his own individual point of view. But nobody would call it a moral value judgment because it would be obviously a judgment based primarily on self-interest (Harsanyi 1977: 631).

Harsanyi (1977: 631-33) then suggests that a moral judgement is one that is made from an impartial point of view. That is, moral judgements are made from a standpoint on which permutations of the identities of the affected parties make no difference to the evaluation of a distribution of goods. In support of this conception of moral judgements, Harsanyi appeals to Adam Smith's (1759/2009) impartial

spectator form of utilitarianism. Indeed, Harsanyi tells us that the prerogative to choose the distribution that maximises average welfare would apply if the person behind the veil were 'an interested outsider rather than a member [of society].' It stands to reason that Harsanyi's justification for our assigning equal weight to each utility function is based on the alleged moral significance of a point of view which Ralf Bader (forthcoming) calls *the point of view of no one in particular* (c.f. Nozick 2001: 75-90; Eddington 1920: 27-39). Bader is careful to note that this point of view presupposes no impersonal moral value over and above the prudential value of each person. But it does require us to broaden the scope of our considerations so as to include the prudential good of others and not just our own prudential good.

The upshot of all this is that Harsanyi's utilitarianism is a contractarian theory only in a very loose sense. If we adopt the point of view of no one in particular, then Harsanyi is correct that it is in our best interests to follow average utilitarianism. But the reason to adopt this point of view is a moral and not self-interested reason. In contrast to Harsanyi, Gogoll and Müller try to ground their utilitarian decision-procedure in the rational self-interest of each road-user without reference to the point of view that morality requires us to adopt. But it is not in the rational self-interest of each road-user to adopt an average utilitarian decision-procedure. This is clear from their comment that 'an interesting question that arises from this line of argument would be whether [a mandatory utilitarian decision-procedure] would incentivize people to car-share to minimize their risk of being targeted' (2017: 697f). The answer to this question is Yes. People would be incentivised to minimise the probability of being targeted by travelling in groups. It is only in everyone's best interests to adopt the average utilitarian algorithm if everyone is indifferent about which utility function corresponds to them. But there are no self-interested reasons to adopt this point of view. Hence it does not follow on contractarian grounds that each of us has reason to want AVs to coordinate to maximise average utility.

1.2.4. Contractualism

I now turn to contractualism. It is helpful to distinguish between contractualism in a broad sense and contractualism in a narrow sense (Ashford and Mulgan 2018). What contractualism refers to in the broad sense is a family of theories that derive the content of morality or the reason to be moral from facts about which principles persons with equal moral status would agree to act in accordance with. This view is different from contractarianism. Contractarianism is about what moral principles

it is rational for people to agree to given their self-interested aims. Contractualism is concerned about what is justifiable to people given their equal moral status.

On the other hand, contractualism in the narrow sense refers to the account of contractualism developed by T.M. Scanlon (1998). According to Scanlon, 'An act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement' (1998: 153). Here an individual has grounds to reasonably reject a moral principle if it gives undue weight to moral claims that are personal to them, such as claims about their welfare, their being treated fairly, or their being treated with respect. I single out Scanlon's contractualism here as it is the best-known and best-defended form of the view.

Unfortunately, Scanlon's contractualism has for the most part been overlooked in the dispute over the moral design problem. This is unfortunate because Scanlon's theory captures much of what is important about morality. Recently, Katherine Evans et al. (ms) have developed a view that is in some respects similar to Scanlon's. This view holds that individuals have moral claims to be treated in one way or another by AVs in collisions. Thus in a collision the AV must balance the competing claims of the different affected parties. One important feature of Evans et al.'s view is that what morality requires is grounded in the moral claims of individuals and not groups. Thus their view satisfies what Scanlon calls the *individualist restriction*, which on Derek Parfit's (2011b: 193) formulation holds that 'in rejecting some moral principle, we must appeal to this principle's implications only for ourselves and other *single* people' (c.f. Scanlon 1998: 229). Evans et al. also maintain that the claims of individuals need not be grounded in facts about their welfare. This renders their position distinct from the competing claims view, on which the strength of my claim to a benefit depends on both how much I stand to gain from that benefit, and also my level of welfare relative to other people who would lose out should I receive that benefit (Otsuka and Voorhoeve 2009; Voorhoeve and Fleurbaey 2012).

In the respects that I have mentioned, Evans et al.'s solution to the moral design problem is Scanlonian in nature. However, their view leaves open how it is that the competing claims of different affected parties are to be traded-off against each other. Evans et al. suggest that AVs could use different 'moral profiles', or algorithms that determine how the claims in a given collision are balanced. In order to fully capture Scanlon's contractualism, I suspect that Evans et al. would need to develop a 'moral profile' that adhered to principles that no one could reasonably reject. But the task of settling on a particular 'moral profile' falls outside the scope of their paper. Their principal aim was to develop a mathematical model of AV decision-making that

could accommodate different conceptions of morality. I do think, however, that their model provides a good basis for developing a Scanlonian collision algorithm.

There is another contractualist answer to the moral design problem. Derek Leben (2017) has developed a collision algorithm that he takes to be based on the contractualist theory of justice proposed by John Rawls (1971, 2001). According to Leben, AVs ought to allocate harm in collisions in accordance with what the affected parties would rationally consent to under fair bargaining conditions. To capture these fair bargaining conditions, Leben develops a veil of ignorance that shares some features with Rawls' veil of ignorance. The parties know the survival probabilities of each affected party conditional on each act available to the AV. But the parties do not know which probabilities correspond to *them*; and the parties do not know how many people correspond to each survival probability (i.e. if two people have the same survival probability, then this is registered as a single survival probability). Leben argues that under these conditions, the parties would agree for the AV to choose the act which maximises the minimum survival probability. Because I provide an extensive discussion of Leben's answer to the moral design problem in Chapter 4, I shall not critique his view here (see also Keeling 2018b).

1.2.5. Deontology

The last of the usual suspects is deontology. Roughly, deontological ethics refers to a family of nonconsequentialist moral theories. What makes a moral theory consequentialist is that it holds that facts about what is morally right obtain solely in virtue of facts about the goodness of states of affairs (Parfit 2011a: 373; Broome 1991: 10-16). Deontological theories are non-consequentialist in the sense that they hold that the right is prior to the good (c.f. Rawls 1971: 30). There have been two notable attempts to provide deontological answers to the moral design problem.

First, Gurney (2016: 217-22) outlines a deontological solution based on Kant's ethics. Kantian ethics is an agent-centred deontological theory, in the sense that it holds that each of us has a set of moral obligations that ground agent-relative reasons to act (Scheffler 1988). Kant evades short summaries. But the basic idea of Kant's ethics is the enlightenment ideal that humans are creatures who can act on principle; and that unlike animals, humans need not be slaves to instinctive drives (c.f. Wood 1999: 331-332). The central tenant of Kant's ethics is that we ought to take the *categorical imperative* as our supreme principle of practical reason. Kant offers multiple formulations of the categorical imperative in his *Groundwork*; the

best known is that we ought to act only in accordance with maxims (i.e. plans of action) that we can rationally will to be universal law (4: 421). To illustrate, suppose I plan to break a promise because it is expedient for me to do so. Then if it were a universal law that each of us could break promises as and when it is expedient to do so, then the institution of promise-keeping would collapse as nobody would take promises to restrict the future conduct of the promiser (c.f. Scanlon 2011: 118).

Because AVs are not moral agents in the Kantian sense, Gurney argues that a Kantian approach to the moral design problem consists in the designers of AVs adopting a Kantian approach to the development of collision algorithms. Gurney's explanation of what a Kantian approach amounts to boils down to a deontological side constraint against killing. This means a moral requirement for AVs *never* to kill. I do not think that this is a plausible interpretation of Kant. For example, Gurney argues that Kantian ethics delivers no verdict about the following case:

> *Tunnel Problem:* 'An autonomous vehicle is travelling along a single lane mountain road that is fast approaching a narrow tunnel. Just before the car enters the tunnel, a child attempts to run across the road but trips in the center of the lane, effectively blocking the entrance to the tunnel. The car has but two options: hit and kill the child, or swerve into the wall on either side of the tunnel, thus killing its operator. How should the car react?' (Gurney 2016: 202).

According to Gurney, 'CI(1) prohibits killing; in either situation, someone dies. Thus, Kantian ethics fails to provide a rule of action' (2016: 221). Here 'CI(1)' refers to the first formulation of Kant's categorical imperative, i.e. agents should act only in accordance with maxims that they can rationally will to be universal law. This strikes me as a fundamental misunderstanding of Kant's categorical imperative.

In addressing the Tunnel Problem, it is helpful to take another formulation of the categorical imperative, the *Formula of Humanity*, according to which we must treat all rational beings as ends in themselves and not merely as means to our ends (G 4:429). Roughly, I treat someone as a means when I use them to achieve some aim. I treat someone *merely* as a means if I use them in whichever way best serves my aims, without accounting for their moral claims against my performing some act (c.f. Parfit 2011a: 212-3). If the AV swerves to avoid the child, it is a foreseen outcome that the passenger will die. Hence in programming the AV to swerve, the AV's designer treats the passenger as a means to saving the child's life. On what Parfit (2011a: 221) calls the *standard view*, 'if we harm people, without their consent, as a means of achieving some aim, we thereby treat these people merely as a means, in a way that makes our act wrong.' So, if the passenger has not given their consent

in advance to be sacrificed in cases like these, it is morally wrong on the standard Kantian view to programme the AV to swerve. Notice that the same reasoning does not hold for the child. The passenger has set the AV to drive to some destination. That a child should fall such that the AV is unable to stop is an unforeseen and improbable event. Hence in hitting the child the AV is not using the child merely as a means to achieve its goal of getting to its destination. The child's death is an accident. But it is not permissible on Kantian grounds to sacrifice the passenger's life to save the child's life. That is one Kantian solution to the Tunnel Problem.

There is a second deontological answer to the moral design problem. Dietmar Hübner and Lucy White (2018) argue that there is a morally salient distinction between persons that are *involved* in a collision and persons that are *uninvolved*. The basic idea is that a person is involved in a collision if the person would be harmed were the AV to take no additional action. For example, imagine that the AV can avoid hitting a pedestrian lying in the road only if it swerves into oncoming traffic. The pedestrian here is involved in the collision: the pedestrian would be harmed unless the AV took action to avoid harming them. In contrast, a person is uninvolved if they would be harmed only if the AV were to perform an additional action. For example, if the AV could avoid a collision with another vehicle only if it swerved onto the pavement and hit a pedestrian. Unfortunately, Hübner and White are non-committal about which moral principle best explains the moral relevance of the involved/uninvolved distinction. Their aim was simply to "introduce these [deontological] elements into the conversation, as [they] are convinced that some sort of distinction between 'involved' and 'uninvolved' must be incorporated into an ethically adequate crash algorithm' (2018: 694). I think that Hübner and White are on the right lines. In Chapter 5, I reject the involved/uninvolved distinction. But the theory I develop is broadly in keeping with Hübner and White's approach.

1.3. The Moral Design Problem Continued

I have discussed some answers to the moral design problem that correspond to first-order normative ethical theories. I shall now discuss three more answers. These answers do not conform to the standard moral theories.

1.3.1. Customisable Ethics Algorithms

The first view holds that AVs ought to have collision algorithms that reflect the moral preferences of their passengers. There are two accounts of this view.

First, Jason Millar (2014) argues that AVs ought to act as 'proxies' for the moral beliefs of their passengers. The bulk of Millar's discussion is about novel medical technologies that make ethically charged decisions. For example, he considers an internal cardiac defibrillator that a patient might wear to monitor for abnormal heart rhythms and deliver an electric shock if the patient's heart arrests. He argues that medical technologies like these ought to reflect the moral beliefs of the patients that are using them. The defibrillator ought to resuscitate the patient only if they have a preference for being resuscitated. The rationale behind Millar's view in the medical case is that there are no objectively correct answers about how the patient ought to be treated beyond what the patient prefers. Millar applies the same point to AVs. He argues that AVs ought to act as 'moral proxies' for their passengers because ethical dilemmas that AVs might encounter have no objective solutions.

This argument strikes me as deeply confused. Medical ethicists have for the most part rejected a paternalistic model of medicine according to which doctors ought to select optimal treatments for patients based on what they consider to be best for the patient. The prevailing view is that patients ought to reach a decision about which treatment is best after reflecting on the best available medical evidence, and evaluating the different treatment options in light of their beliefs and values (Beauchamp and Childress 1994; Faden and Beauchamp 1986). The dominant view places autonomous decision-making on the part of the patient as the aim of clinical deliberation. The role of the doctor is to guide the patient through the decision. It is not to make the decision on behalf of the patient. Accordingly, in medicine it makes sense for treatment technologies to align with the values of the patient.

The same point has no straightforward application to AVs. One morally salient difference between AV collisions and medical decision-making is that in medicine the patient is standardly the only person whose life is at stake in the decision. But in collisions, the AV is required to balance risks of death or harm across multiple persons whose interests are in conflict. The problem with paternalistic medicine is that doctors do not respect the patient as an autonomous decision-maker if they impose treatments upon patients that patients do not want. But allowing an AV's passenger to dictate how the AV behaves in a scenario where multiple lives are at stake assigns inappropriate weight to the autonomous decisions of the passenger. I

do not understand why the passenger's preferences should be given full weight, at the expense of giving no weight to the preferences of the other affected parties.

You might think that Millar's view can be made plausible by restricting the degree to which the AV is partial to the passenger. If the passenger is given *some* freedom to decide how partial the AV is towards them within plausible bounds, then the view is more attractive. This move has a serious cost for Millar. The problem is that Millar would then owe an account of the moral principles that underpin the restrictions on the AV's partiality. But because Millar's view is predicated on the idea that appeal to moral principles undermines the autonomy of the AV's passenger, he is unable to make this move without rendering himself vulnerable to the same criticism. Furthermore, the view that Millar defends renders itself explanatorily redundant if it is couched in terms of a more basic moral theory. In short, if Millar's view works only if the passenger's preferences are restricted so as to conform to independent moral standards, then what is doing the explanatory heavy lifting is the independent standards as opposed to Millar's view. This is the central insight of Plato's *Euthyphro*. I shall discuss this point further in Chapter 5.

Giuseppe Contissa et al. (2017) provide a better argument for the view that AVs should have collision algorithms that reflect their passengers' moral outlooks. They argue that AVs should be equipped with a customisable 'ethics knob' that allows the passenger to input the degree to which the AV prioritises their safety over the safety of other road-users. Their argument for this view is that marketing AVs with a customisable ethics setting might help to promote the general acceptance of AVs.

As far as I can tell, Contissa et al.'s argument is most plausible if we accept a utilitarian criterion of rightness. If what matters is saving lives, then there may be good reason to produce AVs that have a customisable ethics setting. As we saw in the discussion of the self-interest theory, AVs that act in their passengers' best-interests are not guaranteed to produce the best outcomes for their passengers. But Contissa et al. are free to maintain that in the long run more lives will be saved through the general adoption of AVs than would be saved if AVs had collision algorithms that disincentivised people from adopting AVs. This is not a particularly strong argument. Regulators could mandate the use of AVs and mandate that in collisions AVs coordinate to produce optimific collision outcomes (Gogoll and Müller 2017). This would save the most lives. Hence the case for Contissa et al.'s 'ethics knob' rests on ignoring certain policy options. Thus I am hesitant to accept that Contissa et al.'s view is the best option under a utilitarian criterion of rightness. In principle, Contissa et al. could also appeal to a criterion of rightness that puts

individual autonomy as the central moral concept. But, as I explained above in the discussion of Millar's view, the dialectical costs of adopting such a view are high.

1.3.2. Political and Legal Approaches

Johannes Himmelreich (2018: 676) has argued that 'insofar as we value the moral diversity of our political community, it should be recognized that [AVs] pose primarily a political problem, not a moral one.' What Himmelreich is suggesting is that moral theory is an inappropriate tool for regulating AV behaviour *because* different stakeholders hold conflicting views about the morally right way for AVs to behave in collisions/normal driving. Himmelreich's view is that our starting point ought to be pluralism. Given that people have reasonable disagreements about the morality of AVs, the principles that regulate AV behaviour ought to reflect a *compromise* between the different views that citizens hold about how AVs ought to behave. Himmelreich claims that reflecting on what morality requires in particular cases, as is standard in moral philosophy, aims 'to elicit an individual's decision.' But these 'cases make no room for [value] pluralism' (2018: 676).

Similarly, Filippo Santoni de Sio (2017) has defended a legal-philosophical approach to the moral design problem (c.f. Keeling 2018a; Coca-Vila 2018). The starting point for this view is much the same as Himmelreich's approach. According to Santoni de Sio, there is widespread disagreement about the moral principles which correctly describe what is morally required in AV collisions. Santoni de Sio suggests that we can appeal to the criminal law to help us to construct a plausible solution to the moral design problem, as the criminal law often contains pragmatic solutions to moral questions about which people disagree. I believe that these two positions are among the most interesting views that have been proposed so far. I discuss Himmelreich's view in detail in Chapter 2 and Santoni de Sio's position in Chapter 3. I will not discuss these views here in order to avoid repetition.

1.3.3. Empirical Approaches

I shall end this treatment of the moral design problem with a discussion of the empirical ethical approaches that have been advanced (c.f. Keeling 2017). There are at this point several examples of empirical scholars investigating the moral design problem. For example, Bonnefon et al. (2016) conducted a series of online surveys

in which participants were asked to judge what the AV morally ought to do in a set of dilemmatic collision scenarios. The sorts of moral questions at issue here include 'should the AV swerve to avoid a pedestrian and kill its driver in the process?' and 'is the AV permitted to kill one pedestrian to save a larger group of pedestrians?'. Bonnefon et al.'s motivation for performing this empirical research was to help AV manufacturers select an AV ethics policy that satisfies three objectives: 'being consistent, not causing public outrage, and not discouraging buyers' (2016: 1573). Insofar as this was Bonnefon et al.'s aim, I am hesitant to say that Bonnefon et al.'s research is an attempt to use empirical methods to solve the moral design problem. Instead the point of the research is to aid empirically-informed policymaking.

I shall flag a couple of problems with Bonnefon et al.'s empirical study. First, the inferences made about the ethical views of participants are invalid. Consider a case in which the AV can either crash into a group of pedestrians, or swerve to avoid these pedestrians but kill one other pedestrian in the process. Bonnefon et al. claim that 'the most common moral attitude is that the AV should swerve'; and that 'this would fit a utilitarian doctrine, according to which the moral course of action is to minimise casualties' (2016: 1573). Here caution is required. Maximising act utilitarianism is not the only moral theory that implies saving the greater number. The participants might reason in accordance with a non-consequentialist argument for saving the greater number (e.g. Scanlon 1998: 232-33; Kamm 1993: 114-19). Hence we cannot infer that participants in the study have utilitarian preferences from the fact that they preferred an option that is consistent with utilitarianism.

Second, Guy Kahane et al. (2015) conducted an empirical study which found no association between making saving-the-greater-number judgements in dilemmas and possessing characteristically utilitarian traits such as impartial concern for others and self-sacrifice. Hence there is good independent reason to think that the apparent utilitarian judgements among Bonnefon et al.'s participants may not reflect moral attitudes that are consistent with utilitarianism. So, insofar as the aim of Bonnefon et al.'s research was to provide policy-relevant data on people's moral beliefs, it is unclear that their methodology is sufficient to achieve this aim.

One other empirical approach is that of Leon Sütfeld et al. (2017). These people conducted a study in which participants were presented with collision scenarios in a virtual reality simulation (c.f. Bergmann et al. 2018; Faulhaber et al. 2019; Keeling 2017; Nezami et al. 2020; Sütfeld et al. 2018). The aim was to extract information about the moral preferences of the participants by gathering data about the participants' responses to different kinds of collisions. These data are intended to serve as the basis for AV decision-making algorithms in collisions. The rationale

behind this empirical research is somewhat perplexing to me (c.f. Keeling 2017). Sütfeld et al. argue that 'there is no ground truth in our ethical intuitions that holds irrespective of context,' such that there is good reason to conduct empirical studies on human decisions so that AVs can replicate our context-sensitive judgements.

Sütfeld et al.'s emphasis on the context-sensitivity of human moral judgements led me initially to assume that they were committed to moral particularism (Keeling 2017: 1). That is, the view that there are no general moral principles (Dancy 2004). However, Sütfeld et al. (2018: 2) responded that 'the experimental data presented and the conclusions based on it are [...] not dependent on a specific position in the views on particularism vs. generalism, but independent of this controversy.' I agree that the data gathered are neutral on the truth or falsity of moral particularism. Though some have argued that empirical data may bear on disputes in first-order normative ethics and in meta-ethics, I do not accept this position (c.f. Joyce 2008; Greene 2001, 2008; see also Berker 2009). I also agree that the conclusions drawn from the data about how humans in fact behave in collisions are independent of the truth or falsity of moral particularism. What I did not understand, and still do not understand, is how these data are relevant to the moral design problem. That people behave in certain ways in collisions does not have any bearing on how AVs ought to behave in collisions. One question is descriptive. The other is normative.

1.4. The Blame Problem

What I am calling the *blame problem* is the problem of who is morally responsible for harm or property damage caused by AVs in collisions. Much less has been said about the blame problem in contrast to the moral design problem.[7]

Why does the blame problem matter? First, our judgements about who is blameworthy for harm that AVs cause in collisions may have implications for our answer to the moral design problem. Indeed, insofar as I give attention to the blame problem in this thesis, it is in relation to the moral design problem. One other reason to care about the blame problem is that as a matter of justice it might be appropriate to hold certain people accountable when AVs cause harm or damage to property. That is to say that when an AV crashes into my front room it may not be

[7] There is a related legal question about who is liable for harm or property damage caused by AVs in criminal and civil law. There have been several interesting treatments of this question. See, for example, Boeglin (2015), Douma and Palodichuk (2012), Duffy and Hopkins (2013), Geistfeld (2017), Gurney (2015), Marchant and Lindor (2012), and Smith (2013). Unfortunately, this legal question falls outside the scope of this thesis.

satisfactory from the point of view of justice to be told that no person owes me compensation because the damage was caused by a robot. Presumably, it is possible for people to be *wronged* by actions performed by AVs; and thus there must be some person or group of people who take ownership of that wrongdoing.

But who is to blame? One view that I shall set aside from the outset is that AVs are themselves morally responsible or partially morally responsible for their actions (c.f. Sullins 2006; Hanson 2009). Raul Halki and Pekka Mäkelä (2019: 265) claim that although there is a sense in which robots such as AVs have autonomy, the kind of autonomy at issue is 'too weak to justify attributions of responsibility.' Recently, Sven Nyholm (2018c) has advanced an argument to this effect. Nyholm argues that the level of agency that AVs have is comparable to that of children, in that both AVs and children are capable of navigating their environments based on internal models of the world, but are limited to setting and pursuing domain-specific goals. Thus whilst AVs have limited autonomy to plan routes and set distances between themselves and other vehicles, the ultimate goal of getting to a destination safely is determined by other agents such as passengers and manufacturers. Children are similar in having limited autonomy to set and pursue domain-specific goals. But these goals are set in the broader context of ultimate goals which are determined by parents or carers. Accordingly, Nyholm argues that AV agency is insufficient for moral responsibility in much the same way as the agency that children have.

I think that Nyholm is correct. I appreciate that AVs are moral entities in the sense that these robots make decisions that are morally evaluable (Johnson 2006). But AVs are collaborative agents, and their ultimate goals are set by humans acting individually or in groups (Nyholm 2018c: 1209-1216; Halki and Mäkelä 2019: 272). I find existing attempts to push back on this position utterly unconvincing. For example, Allan Hanson (2009) argues that insofar as technologies extend human agency, at least some responsibility for the relevant act transfers to the technology. Because this position admits obvious counterexamples, e.g. that guns are in part responsible for killing, Hanson employs a weakened notion of responsibility that is divorced from related notions such as moral desert and intention. On this view, guns, cars, and bicycles, can share responsibility for actions. Whilst I am open to the possibility that Hanson's weakened notion of responsibility serves a conceptual role in some domains, I am afraid I do not see its relevance to moral philosophy.

So who is responsible? One distinction that is helpful to draw at this point is whether an AV caused harm or property damage as part of its normal functioning, or whether the harm or property damage was the result of a malfunction. Suppose that in one case the AV crashes into a wall to avoid a pedestrian who stepped out

into the road unexpectedly. Then suppose that in another case the AV crashed into a wall because a hacker uploaded a virus to the AV's computer which interfered with the AV's object classification software. In the latter case, it seems clear to me that the person responsible for the property damage is the hacker. But in the former case, the obvious candidates for moral responsibility are the AV's passenger or its manufacturer; although the responsibility here is likely to be diminished (perhaps entirely) because there was a positive moral reason to crash into the wall. These cases suggest that the blame problem is unlikely to have a one-size-fits-all solution. The responsible agent will likely vary with contextual features of the collision; and the degree of responsibility is also unlikely to be invariant across contexts.

When might an AV manufacturer be responsible? Alexander Hevelke and Julian Nida-Rümelin (2015) have advanced an argument against holding manufacturers morally responsible for AV actions *simpliciter*. Their argument is that holding AV manufacturers responsible is likely to deter the production of AVs, and this would in the long run result in a large number of preventable road-traffic deaths (c.f. Goodall 2017; Fleetwood 2017; Lin 2016). I find this point unpersuasive. I agree with Nyholm (2018b: 3) that the argument 'does not settle the question of whether it is just or fair to hold car manufacturers responsible for harms or deaths their cars might cause.' What matters for the blame problem is not who it is expedient to blame but who is in fact to blame. Furthermore, it seems clear to me that there exist circumstances in which AV manufacturers bear at least some moral responsibility for the AV's actions. This is true, for example, in cases where the AV causes harm or property damage because of a software bug that we could reasonably expect the manufacturers to detect as part of the quality assurance process. It might also be true in cases where the manufacturer cut corners to save costs in the design process, and in doing so exposed passengers or road-users to an undue risk of harm.

What about passengers? Hevelke and Nida-Rümelin (2015) consider two ways to establish passengers in AVs as morally responsible for harm or property damage caused. The first is to establish that AV passengers have a duty to intervene in collisions, i.e. a duty to take control of the AV and determine its behaviour. Hevelke and Nida-Rümelin set aside this view on the grounds that it is unreasonable to expect AV passengers to remain alert and ready to take control of the AV over long periods of time. The second approach is to argue that AV passengers are responsible as a matter of strict liability. This argument holds that because AV passengers knowingly impose a risk on road-users/property by making AV journeys, these passengers are responsible for any harm or damage to property that occurs as a

consequence of the AV's actions. Hevelke and Nida-Rümelin reject this argument on the grounds that it renders passenger responsibility a matter of luck.

These arguments are too quick. First, whether or not it is reasonable to expect AV passengers to be prepared to take control at all times on the road must be evaluated in light of the risk that AV passengers impose on other road-users. I might decide to drive for ten hours in a manual vehicle, become tired, and then kill a pedestrian by mistake. That I cannot pay attention for ten hours straight does not diminish my responsibility for killing the pedestrian. Presumably, morality requires that I drive only insofar as I am able to do so whilst paying attention. What matters is not whether it is reasonable to demand that I pay attention for ten hours. The question is: Given the risk that my driving imposes on others, is it reasonable to demand that I pay attention when driving? There may be good empirical reason to think that people cannot supervise AVs for extended periods (Ruscio et al. 2015). But given that many instances of driving are morally optional, it does not follow that it is unreasonable to demand that people pay attention when operating AVs.

However, my concern in this thesis is with Level 5 AVs, and I am assuming that these AVs lack manual controls. So, I shall bracket this issue. As it stands, I am also unconvinced by Hevelke and Nida-Rümelin's point about moral luck. This point has been made before in the ethics of killing by Helen Frowe (2014: 82) in response to Jeff McMahan's (2009: 165) view that conscientious drivers who are driving for morally optional reasons bear responsibility for harm caused in fluke collisions, as they knowingly impose a risk of harm on other road-users. According to Frowe, it is unreasonable to hold conscientious drivers responsible for harm caused in fluke collisions because these collisions are so improbable, and it is inappropriate to hold people responsible for events that they are justified in believing will not obtain. I argue in Chapter 5 that neither Frowe nor McMahan is entirely right here. But it is not in the least bit obvious to me that the possibility of *moral luck* playing a role in an account of moral responsibility renders that account false. If Hevelke and Nida-Rümelin are right, then it is also true that actual harm and attempted harm attract the same degree of moral responsibility; and this is certainly not obvious.

1.5. The Risk-Imposition Problem

The last problem that I shall discuss is the *risk-imposition problem*. The AV has a prudential goal of getting to its destination in good time. But the faster it drives, the more risk of harm it imposes on road-users. How does the AV's moral goal of

road-user safety trade-off against its prudential goal of time-efficiency? What is the morally right amount of caution for AVs to exercise when uncertain about what kinds of objects are in their environments and how those objects will behave?

The risk-imposition problem is about the morality of normal driving; it is not about collisions (c.f. Himmelreich 2018: 678-81). Several people have articulated the risk-imposition problem and defended its importance (Goodall 2016, 2019; Himmelreich 2018; Keeling et al. 2019; Nyholm and Smids 2016). But there has been little, if any, systematic treatment of the problem in the AV ethics literature. The problem of risk-imposition has been discussed elsewhere in moral philosophy. I intend to take these discussions as a starting point. In what follows, I shall survey the positions in this dispute using the morality of drunk driving as a case study.

Most people think that drunk driving is morally wrong even if it harms nobody (c.f. Finkelstein 2002; Oberdiek 2009, 2012; Steinbock 1985). Presumably, the best explanation for this judgement is that drunk driving imposes a more serious risk of harm on road-users than does sober driving. There have been a handful of attempts to elucidate what is morally wrong about risk-imposition due to drunk driving. We can learn from this dispute some important considerations for our own problem.

Claire Finkelstein (2002) argues that imposing risks of harm itself constitutes a form of harm. Hence drunk drivers that do not inflict material harms upon people still harm other road-users in virtue of exposing them to a risk of harm. Finkelstein understands harm as a setback to a person's interests; and a benefit as something that promotes someone's interests. Finkelstein argues that it is possible to promote or set back a person's interests by giving them a chance of a material benefit or harm. If I buy you a lottery ticket, then I benefit you even if the lottery ticket does not win. What is beneficial is that I have given you a small chance of winning a significant amount of money. Finkelstein calls this a *risk benefit*. On the other hand, if I force you to play Russian Roulette, and impose a one in six chance of death upon you, then I set back your interests. This is true, Finkelstein argues, even if you survive. What I have done counts as a *risk harm*. Finkelstein claims that drunk drivers impose risk harms on other road-users. Because risk harms are harms, drunk driving counts as harming people even if no material harm obtains.

Doug Husak (1994) presents an interesting problem for Finkelstein. Husak points out that the probability that a drunk driver will kill someone on any given journey is not significantly higher than the probability that a sober driver will kill someone. The probabilities in both cases are extremely close to zero; and the absolute difference is not that great (c.f. Ori 2014, 2015). Presumably, what matters

for risk harms is the absolute as opposed to relative risk of a material harm. Hence it is unclear how we can sustain the view that drunk driving is wrong *because* it imposes an unjustifiable risk of harm on road-users; but sober driving is permissible even though the probability of a sober driver harming someone is not much lower.

Husak's argument invites a range of responses. First, we might bite the bullet and claim that the moral harms of sober driving have been greatly underestimated. McMahan (2009: 117) emphasises in his account of permissible killing that driving for morally optional reasons imposes a small risk of serious harm on road-users; and that as a consequence even conscientious drivers ought to bear the costs in unavoidable collisions. Second, we might argue that it is misplaced to locate the morally relevant sense of risk in the probabilities. Instead we might opt for a modal notion of risk. For example, we might claim that drunk drivers *could easily* kill pedestrians, in the sense that there are close possible worlds in which situations arise where the drunk driver cannot safely control the vehicle to avoid colliding with a pedestrian. But for sober drivers, the possible worlds in which they are presented with scenarios in which they cannot safely control their vehicle are much more distant (c.f. Pritchard 2015, 2016; Ebert, Smith and Durbach 2019).

Third, we might look for other explanations for what is morally problematic about drunk driving. David Oberdiek (2009, 2012) agrees with Finkelstein that risk-imposition constitutes a form of harm. But Oberdiek differs in his explanation for the harm. According to Oberdiek, risk-impositions are morally wrong because they diminish the autonomy of the victim. The notion of autonomy that Oberdiek has in mind is one on which it matters morally that people are free to choose their own life plans, and that in doing so they have a reasonably large set of acceptable options to choose between. On Oberdiek's view, risk-impositions are problematic insofar as they reduce the number of acceptable options available to people. What drunk drivers do that is morally wrong is narrow the safe options available to road-users in terms of where they can physically go whilst avoiding material harms.

One final approach is to abandon the idea that risk-imposition constitutes a form of *harm*, and instead claim that people have a *right* not to have risks of harm imposed upon them (McCarthy 1997). This view has been widely criticised. The problem is that it faces what Madeleine Hayenhjelm and Jonathan Wolff (2011: 37) call the *paralysis problem*, according to which a right against risk-imposition would rule-out almost all acts that are available to us at any given time. There is always some risk that an act might have some unintended consequence. The boiler might explode when I turn the heating up; the wheels might fall off the car as I speed up to join

the motorway. Hence a right not to have risks of harm imposed upon us would leave us paralysed in the sense that it would be impossible to do *anything* without violating people's rights (Fried 1970: 192-3; Holm 2016: 917-8; Kagan 1989: 87-88; Nozick 1974: 73-78; Schroeder 1986: 527). Hence the rights view is problematic for at least the reason that it implies that almost all of us act wrongly all the time.

There are responses to the paralysis problem. One option is to claim that rights are best understood as *pro tanto* rights, such that in cases where there is a positive moral reason to impose a risk on someone, their rights are not violated (McCarthy 1997; Holm 2016: 924-29). That is, the right not to have risks of harm imposed upon us can be outweighed by countervailing considerations. One other option is to hold that rights against risk-imposition can be waived through *consent*, and that minor acts of risk-imposition receive tacit consent, or that acts of risk-imposition can be *compensated* for after the fact (c.f. Nozick 1974: 73-78). This is not the place to explore these responses in detail. I shall return to the problem of risk-imposition in Chapter 6 when I develop my own solution to the risk-imposition problem.

1.6. Plan of the Thesis

This thesis is a collection of papers. It is not a monograph. Together these papers are intended to support the deontological position that I defend. Here I shall explain how the different chapters work together to support my conclusion.

In Chapter 2, I discuss methodology. Throughout the thesis, I use *trolley cases* to illustrate points and to argue for and against different views. The relevance of these imagined cases to the moral design problem has been challenged (Goodall 2016; Himmelreich 2018; Nyholm and Smids 2016). Because I do not think that the arguments against trolley cases are persuasive, and because I rely on these cases to defend certain claims, it is right that I respond to this challenge. I shall tackle the four strongest arguments against the use of trolley cases; and in doing so, develop a positive account of how trolley cases might inform the ethics of AVs.

In Chapters 3 and 4, I argue against two rival non-consequentialist answers to the moral design problem. In Chapter 3, I tackle Filippo Santoni de Sio's (2017) legal-philosophical approach. I have chosen to examine this view because what it recommends is not dissimilar to what my view recommends. However, Santoni de Sio thinks that he can derive such a view from much weaker foundations. Recall: the rationale behind the legal-philosophical approach is that people disagree about

which moral principles are true, but they could in principle agree to a solution to the moral design problem based on the criminal law. Santoni de Sio then presents an answer to the problem based on the legal doctrine of necessity. I argue that Santoni de Sio's answer to the moral design problem does not achieve its aim of moral agreement as it tacitly appeals to moral principles that many people reject. I then provide a novel formulation of the doctrine of necessity based on two legal judgements, those of Lord Coleridge in *Dudley and Stephens* (1884) and Lord Brooke in *Re A* (2001). My account of the doctrine of necessity appeals to a Restricted Pareto Principle, which details the *conditions under which* a killing is permissible, as opposed to *that in virtue of which* a killing is permissible. I argue that the Restricted Pareto Principle is agreeable by the lights of three of our best moral theories. But I reject the principle as it applies only in a small subset of cases (and any attempt to relax the principle renders it disagreeable by the lights of at least one theory).

In Chapter 4, I critique Leben's (2017) 'Rawlsian' solution to the moral design problem. I have chosen to discuss Leben's view because it is the best-developed non-consequentialist position in the literature; and thus provides a serious rival to the deontological account that I defend. I argue that Leben's algorithm is based on a misreading of Rawls, such that Rawls' account of justice offers no support to Leben's algorithm. Hence Leben owes an independent argument for his view. I then present three obstacles to Leben providing such a justification. The first is that under plausible assumptions about rational preferences, a rational agent could not consent to the prescriptions of Leben's algorithm in certain collisions. Second, Leben's algorithm gives undue weight to the moral claims of the worst-off. Third, given Leben's assumptions about what matters morally, there is a better algorithm that ought to be used in place of his 'Rawlsian' algorithm. However, I argue that we ought to reject this improved algorithm because it is insensitive to the positions of the affected parties in collisions, e.g. who is morally responsible for the crash.

In Chapter 5, I present my own solution to the moral design problem. Because I take facts about moral responsibility to be relevant to what the AV ought to do, I also go some way towards answering the blame problem. My view is that the AV is morally permitted to kill or harm a road-user if, and only if, and because, its passengers are permitted to kill or harm that road-user in self-defence. I focus on providing answers to practical ethical questions presented by plausible road-traffic scenarios. How much risk of harm is the AV passenger morally required to undertake to avoid killing or harming a jaywalking pedestrian? To what extent is the AV required to prioritise the safety of pedestrians in loss-of-control cases? How should the AV balance the competing interests of its passenger and other road-

users if another road-user blocks the AV's escape route from an independent threat? The self-defence view that I defend attempts to answer these questions in line with a rights-based deontological account of killing. I argue that this view does a better job than its rivals at capturing our considered moral judgements; and that the view better coheres with our best theories of killing and harming in other domains.

In Chapter 6, I develop an answer to the risk-imposition problem. This answer is broadly in line with the deontological theory that I develop in Chapter 5. I focus on situations in which the AV is uncertain about the classification of a proximate object, much like the 2018 Tempe Arizona AV collision (NTSB 2018). I develop a picture of how the AV morally ought to balance its prudential goal of *time-efficient driving* against its moral goal of *road-user safety* in these uncertain situations. The account holds that the AV is morally permitted to pursue its prudential goal insofar as its speed and position make it the case that, given the AV's evidence, the AV could not easily kill or harm a pedestrian in a close-by *what if* case. Thus I appeal to a modal notion of risk in place of the standard probabilistic conception of risk. I then consider the implications of the view in practice. I model a mundane road-traffic scenario as a Markov Decision Process (MDP). I then consider how the view I defend might impact our design choices for the model parameters. I focus on the AV's reward function and on the stability of its classification predictions over time.

2. Why Trolley Problems Matter[8]

There is, I think, a great deal of misunderstanding about trolley cases and their relation to the ethics of automated vehicles (AVs). In this chapter, I shall do what I can to put this right. I want to argue that those who defend the use of trolley cases in the ethics of AVs do so for the wrong reasons; and that those who argue against these cases misinterpret the role of trolley cases in moral theorising. I understand that in defending this thesis I put myself at odds with all parties in the dispute. But I hope that careful reflection on the insights from both camps might bring us closer to a consensus on the relevance of trolley cases to ethical disputes about AVs.

I shall not bother with a rigorous definition of trolley cases. Roughly, these are imagined cases in which the AV must choose one of two acts; the consequences of each act are known; each act imposes a distribution of benefits and burdens over at least two parties; and the interests of these parties are jointly unsatisfiable. In one such case, the AV can continue its course and cause the deaths of five pedestrians; or swerve to avoid the pedestrians but kill its passenger in the process (Bonnefon et al. 2016: 1574). Obviously, these cases are called *trolley cases* because they are in certain respects similar to cases in the ethics of killing that involve runaway trolleys (Foot 1967/2002; Kamm 1993, 2007, 2016; Thomson 1976, 2008).

Typically, people argue for the relevance of trolley cases to AV ethics by giving examples of dilemmatic collisions that AVs could in principle face (Lin 2016: 76; Leben 2017: 107; Goodall 2014: 60-61). The argument holds that *because* AVs may face these collisions in practice, we need to know what to do should they arise. On the other hand, people argue for the irrelevance of trolley cases to AV ethics by pointing out that dilemmatic collisions are unrealistic; or that trolley cases ignore morally salient features of collisions (Goodall 2016, 2019; Nyholm and Smids 2016; Himmelreich 2018). What I claim is that the relevance of trolley cases to AV ethics

[8] This chapter is published, in a revised form, as 'Why Trolley Problems Matter for the Ethics of Automated Vehicles' in *Science and Engineering Ethics*, 2020, 26, pp. 293-307.

does not depend on AVs facing dilemmatic collisions in practice; and that it is difficult to make plausible claims about moral principles for the general regulation of AV behaviour without examining what morality requires in trolley cases.

To argue for this view, I shall consider four arguments against the use of trolley cases in AV ethics, and show that my picture of how these cases ought to be used has the resources to weather these criticisms. The arguments that I address are as follows: (i) that trolley cases are not going to happen in practice; (ii) that there is a substantial moral difference between trolley cases and real-world collisions, e.g. that the latter involve uncertain decision-making and the former do not; (iii) that it is impossible for AVs to deliberate in accordance with moral theories based on trolley cases as these cases presuppose a 'top-down' approach to AV engineering; and (iv) that trolley cases provide a moral answer to what is ultimately a political problem. Throughout, I shall refer to the *general moral design problem*, as the problem of determining a principle or set of principles for the general regulation of AV behaviour at all times on the road. Thus the general moral design problem includes both the moral design problem and the risk-imposition problem.

2.1. The Not Going to Happen Argument

2.1.1. The Argument

The simplest arguments are sometimes the most instructive. I start with one such argument. According to what I call the *Not Going to Happen Argument*,

> P1. Trolley cases are relevant to real-world collisions in which harming at least one person is unavoidable, and a choice is required about how to distribute harms between multiple persons whose interests are in conflict.
>
> P2. AVs will not encounter dilemmatic collisions like these.
>
> C. Therefore, trolley cases are irrelevant to the general moral design problem.

There are, of course, more relaxed formulations of this argument. For example, Julian de Freitas et al. (2019) claim that trolley cases are 'vanishingly unlikely on real roads [so that without] evidence that such situations occur […] it is unhelpful to consider them when making AV policies.' Similarly, Noah Goodall (2019) notes that '[the] most common response [to trolley cases is] that trolley problems are avoidable, implausible, rare, and distractions from more productive efforts.'

The Not Going to Happen Argument is deeply mistaken. But there are lessons to be learned from this argument. So, it is important that we understand why its proponents have advanced this criticism of trolley cases. The first treatments of AV ethics tried to motivate the subject on the grounds that AVs could in principle face dilemmatic collisions. These include Goodall's (2014: 60) case involving a bus that swerves in front of the AV such that the AV must decide whether to brake or swerve given different distributions of harm conditional on each act; and Patrick Lin's (2016: 71-2) case involving an AV that can either collide with a schoolbus or swerve off the side of a mountain. The Not Going to Happen Argument offers a plausible response to these attempts to motivate the relevance of trolley cases to AV ethics. If the reason to care about trolley cases is that AVs might face these cases, then the fact that AVs will not face these cases implies that trolley cases do not matter.

But this was not the right way to motivate trolley cases in relation to AVs. In responding to the Not Going to Happen Argument, I shall grant P1 and P2. There is no reason to doubt P1. P2 is hard to assess in the absence of empirical data or good computer simulations about the sorts of collisions that AVs might encounter. But there is some armchair evidence for P2. Johannes Himmelreich (2018: 673-4) argues that given plausible assumptions about AVs, it is impossible for an AV to encounter a trolley-style collision whilst at the same time maintaining the level of control required to make a meaningful ethical decision. Himmelreich's point is that a trolley dilemma is unlikely to arise if the AV is travelling sufficiently slowly to make a meaningful ethical decision; and that in the sorts of high-speed cases where a trolley-style dilemma might arise, it is unlikely that the AV would have sufficient control to decide how to allocate benefits and burdens between the affected parties. Because I do not need to argue against P2, I shall assume Himmelreich is correct.

The problem is that the Not Going to Happen Argument is invalid. This means that it is possible for P1 and P2 to be true and for C to be false. To guarantee the truth of C, a third premise is required. The argument should in fact be:

P1. Trolley cases are relevant to real-world collisions in which harming at least one person is unavoidable, and a choice is required about how to distribute harms between multiple persons whose interests are in conflict.

P2. AVs will not encounter dilemmatic collisions like these.

P3. Cases of type X are relevant to the general moral design problem only if AVs will encounter cases of type X in the real-world.

C. Therefore, trolley cases are irrelevant to the general moral design problem.

This argument is unsound. P3 is false. When used correctly, trolley cases can help us to make inferences about what is morally required in AV driving. This is true even if AVs will not encounter trolley cases in practice. To make a convincing case for this claim, I need to do two things. The first is to make precise the model of morality that underpins trolley cases. The second is to illustrate how in principle trolley cases might inform questions about the morality of AV decision-making.

2.1.2. Modelling Morality

The trolley case method is based on a model of morality. This model is rarely made explicit in applied ethics. Though it is made explicit in meta-ethical disputes, the level of precision with which the model is formulated is far from ideal. If we are to understand the model and its limitations, some amount of mathematics is required. I will start with the intuitive picture. Then I will present the model formally, and use the formal model to provide a general explanation for certain counterexamples that have been raised in the literature. I then provide a short defence of the model.

The trolley case method assumes that morality can be represented using what Selim Berker (2007) calls the *generalised weighing model*. According to this model, the moral status of an act φ is determined on the balance of *pro tanto* reasons. Here a *pro tanto* reason is a reason that has genuine weight but that can in principle be outweighed by countervailing reasons (Broome 2013: 51-62; Urmson 1975: 115). The basic idea is that whether or not I ought to φ depends on the balance of reasons for and against φ-ing, taking into account the strength of each reason. If the reasons for and against φ are on balance positive, then I ought to φ.[9] Will MacAskill (2016) and Samir Okasha (2011) have used voting theory to model normative uncertainty and the balancing of theoretical virtues for theory choice in science. I follow their lead in pitching the generalised weighing model as a weighted majority voting problem. Roughly, the φ-relevant factors *vote* on the moral status of φ, and each factor has a certain share of the vote which corresponds to the strength of the reason grounded in that factor. The model is also formally equivalent to a simple perceptron algorithm in machine learning (c.f. Crama and Hammer 2011: 405-8).

[9] I use 'ought to φ' as synonymous with 'most reason to φ'. Note: I do not have an extra reason to φ in virtue of having most reason to φ. That I have most reason to φ is a claim about the relative strength of the *pro tanto* reasons for and against φ-ing (Dancy 2004: 16).

The model has three parts. These parts correspond to what Berker (2007: 116) calls the *underlying level*, the *contributory level*, and the *overall level*. Roughly, the underlying level concerns the normative factors that are relevant to whether or not I ought to φ. Then the contributory level concerns the normative reasons grounded in each of the φ-relevant factors; and the overall level concerns how those reasons are traded-off against one another to determine the overall moral status of φ.

First, the underlying level concerns the normative factors relevant to whether or not I ought to φ. These factors are empirically-discoverable non-normative facts that ground reasons for or against φ-ing (Kagan 1988: 17-22). Suppose that φ is the act of going to the dentist tomorrow.[10] The normative factors in this case might include (i) whether or not I have a toothache and (ii) whether or not I will have to pay for the appointment. Given n normative factors relevant to φ, we can denote these factors $x_1, x_2, \dots, x_n$, such that $x_i = 1$ if factor i obtains, else $x_i = -1$. Thus if factor 1 is whether or not I have a toothache, then $x_1 = 1$ if I have a toothache, and $x_1 = -1$ if I do not have a toothache. It is helpful to treat all the φ-relevant factors as a single mathematical object called a *vector*. John Broome (1991: 66) tells us that '⌜a⌝ vector is simply a list.' This is a comforting half-truth. For the moment, it is helpful to think of vectors as points in n-dimensional Euclidean space:

$$\mathbf{x} = (x_1, x_2, \dots, x_n)$$

Here $\mathbf{x}$ is the vector that contains all the φ-relevant factors. $\mathbf{x}$ represents a possible *factor-combination*. Each factor-combination can be plotted in a *factor-space*. If there are two φ-relevant factors, the factor space is $\mathbb{R}^2$, and $\mathbf{x}$ can take one of four coordinate values: $(1,1)$, $(1,-1)$, $(-1,1)$, $(-1,-1)$. In the dental case, $(1,1)$ is the state of affairs where I have a toothache and I have to pay for the appointment; $(1,-1)$ is the state of affairs where I have a toothache and I do not have to pay; and so on and so forth. The four possible factor-combinations are the vertices of a two-by-two square around the origin (**Figure 1**). In three-factor cases, the space is $\mathbb{R}^3$, and the eight factor-combinations form the vertices of a two-by-two-by-two cube. The general rule: For n normative factors, the factor-space is $\mathbb{R}^n$, and the 2^n factor-combinations form the vertices of an n-dimensional hypercube around the origin.

[10] This example uses practical reasons as opposed to moral reasons. I think the model is a little more intuitive if approached from the angle of practical reasoning in the first instance.

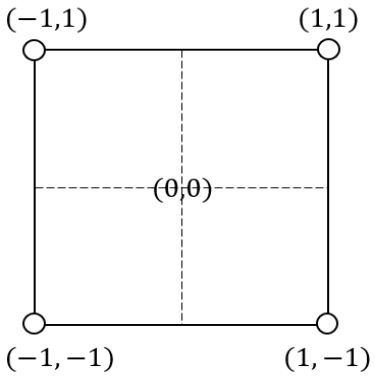

Figure 1: Factor Combinations for Two Factors

I now turn to the contributory level. This level concerns the reasons grounded in each of the normative factors. There are two features of reasons that the model needs to capture. The first is the *valence* or *polarity* of a reason. That is, whether the reason counts in favour of φ-ing or against φ-ing (Dancy 2004: 6; Hooker 2003: 8). That I have a toothache counts in favour of going to the dentist. That I will have to pay counts against. The other feature of reasons that I need to capture is the normative *force*, *strength*, or *weight* of a reason (Berker 2007: 113-4). That I have a toothache does not just count in favour of going to the dentist, it counts in favour to some degree. The same holds in reverse for my having to pay. How do we model the valence and strength of the reasons grounded in the factors x_1 through x_n?

Suppose that each normative factor x_i has a real-valued weight w_i. This number tells us about both the polarity and the strength of the reason grounded in x_i. First, if w_i is positive, then $x_i = 1$ counts in favour of φ-ing. If w_i is negative, then $x_i = 1$ counts against φ-ing. Hence the sign of w_i represents polarity. Second, I shall represent the strength of a reason with the magnitude of w_i. The strength of the reason is determined by the distance between w_i and zero. To keep things simple, I shall represent the weights w_i for the factors x_1 through x_n as a vector:

$$\mathbf{w} = (w_1, w_2, \ldots, w_n)$$

I said earlier that we can think of vectors as points. We can also think of them as arrows. We can plot $\mathbf{w}$ as an arrow in the factor-space that runs from the origin to the point $(w_1, w_2, \ldots, w_n)$. Suppose in the dentist case that my having a toothache has a weight $w_1 = 1$, and that my having to pay has a weight $w_2 = -1$. We can then plot the weight vector as an arrow from the origin to $(1, -1)$, as in **Figure 2**.

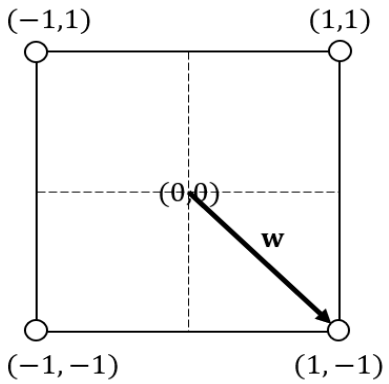

Figure 2: The Weight Vector

Last, I turn to the overall level. This level concerns how the *pro tanto* reasons are balanced to determine what I ought to do. The model holds that there exists a function that takes factor-combinations $\mathbf{x}$ as inputs, and returns 1 or -1 depending on whether or not I ought to φ. Kagan (1988: 14) calls this the *governing function*; and Berker (2007: 120-1) calls it the *combinatorial function* (c.f. Broome 2004a: 37).[11] Trolley cases are based on a special case of the generalised weighing model called the *additive model*. This model holds that the governing function is a weighted sum. That is to say that the governing function adds together each factor multiplied by its weight. I ought to φ if the reasons are on balance positive; else, I ought not φ.

$$f_b(\mathbf{x}) = \sigma\left(\sum_{i=1}^{n} x_i w_i + b\right)$$

Here $\sigma: \mathbb{R} \to \{1, -1\}$ is a threshold that returns -1 if the input is at most zero, and 1 if the input is strictly greater than zero. Thus I ought to φ if $f_b(\mathbf{x}) = 1$; and I ought not φ otherwise. Thus the function captures the idea that I ought to φ if the reasons for and against φ-ing are on balance positive. In two-factor cases, the model works by separating the factor-space into two with a straight line:

$$\mathcal{L} = \{(x, y): w_1 x + w_2 y + b = 0\}$$

I will call this line the *decisional frontier*. Factor-combinations on one side of the frontier are such that $\Sigma w_i x_i + b > 0$. These are the cases in which I ought to φ. Factor-combinations on the other side of the frontier are such that $\Sigma w_i x_i + b \leq 0$. These are the cases in which I ought not φ. The frontier is orthogonal to the weight

[11] The dispute between moral generalists and moral particularists is essentially a dispute about whether this function is finitely expressible (Berker 2007; McNaughton 1996).

vector (see **Box 1** for a proof). The b term shifts the decisional frontier away from the origin, and thus improves the flexibility of the model. **Figure 3** shows the need for the b term in the dentist case. Here I assume for illustrative purposes that the only case where I ought not go to the dentist is where I do not have a toothache and I nevertheless have to pay for the appointment. Consider,

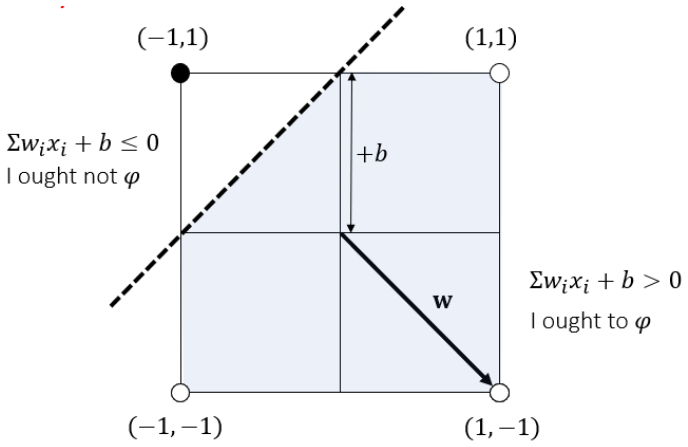

Figure 3: The Need for a Bias Term

The points I have made generalise to cases involving n factors. The factor-space $\mathbb{R}^n$ is divided with the hyperplane $\Sigma w_i x_i + b = 0$. The factor-combinations $\mathbf{x}$ such that $\Sigma w_i x_i + b > 0$ are the cases in which I ought to φ; and the factor-combinations $\mathbf{x}$ such that $\Sigma w_i x_i + b \leq 0$ are the cases in which I ought not φ.

Box 1: Proof that the decisional frontier is orthogonal to w

First, the weighted sum of reasons is equivalent to the dot product of the two vectors $\mathbf{w}$ and $\mathbf{x}$. That is,

$$\sum_{i=1}^{n} w_i x_i = \mathbf{w} \cdot \mathbf{x} = \|\mathbf{w}\|\|\mathbf{x}\| \cos(\theta)$$

Here θ is the angle between $\mathbf{w}$ and $\mathbf{x}$, and $\|\boldsymbol{\mu}\|$ is length of $\boldsymbol{\mu}$, which is $\|\boldsymbol{\mu}\| = \sqrt{\mu_1^2 + \cdots + \mu_n^2}$. Second, we can rearrange the above for θ:

$$\theta = \arccos\left(\frac{\Sigma w_i x_i}{\|\mathbf{w}\|\|\mathbf{x}\|}\right)$$

Third, the decisional frontier is such that $\Sigma w_i x_i = 0$. Hence:

$$\theta = \arccos(0) = 90° \blacksquare$$

That concludes the model. The point of trolley cases is to determine whether particular factors have moral relevance. Typically, trolley cases come in pairs. The aim is to contrast cases c and c', such that the features of these cases are identical bar one. Then the argument goes that if the moral status of the acts φ and φ' in these cases are different, it must be the case that the feature that is permuted has moral relevance (in the sense that it is included as a component in the vector $\mathbf{x}$ that is the input to the governing function for moral rightness). Consider,

> [We] can see why the cases that have been the focus of attention in trolley problem discussions seem artificial and unrealistic. They are specifically constructed, like scientific experiments, to distinguish among and test theories and principles […] because one theory or principle would imply the permissibility of conduct that the other theory or principle would deny. Using our intuitive judgements about which implications for cases are correct helps us decide among, and also revise, theories and principles (Kamm 2016: 13).

That, I hope, clears things up somewhat. But it would be silly of me to pretend that there are no legitimate objections to this model. I shall end by looking at the most serious problem. Shelly Kagan (1998) argued that the method of contrast cases is based on two assumptions. The first is that morally relevant features are *ubiquitous*, i.e. that if a feature is morally relevant in one case, then it is morally relevant in all cases. The other is that the reasons grounded in these features are *strongly separable.*[12] That is, the contribution that each morally relevant property makes to the moral status of an act is independent of the contribution that any other morally relevant property makes; and the same holds for sets of morally relevant properties that together partition the properties that determine the moral status of the act. That morally relevant features interact in such a simple way is questionable.

I do not have a complete response to Kagan. But I can say the following. First, the additive model has quite a bit going for it. Errol Lord and Barry Maguire (2016: 14), for example, write that 'additivity, or something like it, is attractive because it is intuitive, workable, and theoretically straightforward.' The fact that Kagan has identified the *possibility* that the governing function is non-additive is on its own uninteresting. Because the additive model has several virtues, what is required is some independent motivation for a non-additive theory of reasons. This motivation

[12] The strong separability condition is called the Sure Thing Principle by Leonard Savage (1972) and Joint Independence by David Krantz et al. (1971). This condition is important because it facilitates additive representation. John Broome (1991: 70-80) provides a philosophical overview of the significance of the strong separability condition.

has invariably come in the form of *counterexamples* to the additive model. The earliest cases were developed by Jonathan Dancy (1993, 2003, 2004). Consider,

> *Dancy's Book Example:* I borrow a book from you, and then discover that you have stolen it from the library. Normally the fact that I have borrowed a book from you is a reason to return it to you, but in this situation it is not. It isn't that I have *some* reason to return it to you and more reason to put it back in the library. I have no reason at all to return it to you (Dancy 1993: 60).

I am afraid that when it comes to Dancy's counterexamples, I agree with Brad Hooker (2003: 14) that they are 'utterly unpersuasive.' There are multiple ways to respond to Dancy in defence of the additive model. Hooker (2003) and Joseph Raz (2003), for example, argue that Dancy's cases underspecify the grounds of the reasons at issue. Roughly, the reason to return the book is grounded in the complex fact that the book has been borrowed and that it is legitimately owned. In contrast, Roger Crisp (2003) distinguishes between ultimate and non-ultimate reasons. The ultimate reason to return the book is that it is the just thing to do. But it is not just to return the book when it is stolen. Crisp argues that the additive model is true, but that it applies to ultimate reasons as opposed to non-ultimate reasons. I think that Crisp is right here. Dancy has found an apparently non-additive example of reasons accrual because he has misidentified the ultimate reasons in the case.

There are, however, some counterexamples that are more troubling. I shall illustrate with an example from Shyam Nair.[13] This example concerns epistemic reasons for belief. But the same point holds in the moral case. Consider,

> You know that John and Bill are rarely found together—they dislike each other and make it a point to avoid each other. There is a party this week and you are wondering whether John or Bill but not both John and Bill will attend. In this setting finding out John will attend is a reason to believe that John or Bill but not both will attend. Similarly, finding out Bill will attend is also a reason to believe John or Bill but not both will attend (Nair 2016: 59-60).

The problem here is that what I ought to believe depends on an exclusive disjunction of two factors. I ought to φ if exactly one factor obtains, but not if both or neither. Cases like these are troubling. The problem is that XOR, the Boolean function that characterises the exclusive disjunction, is not linearly separable. That a threshold sum cannot represent the XOR function is a well-known problem in

[13] For similar cases see Horty (2012: 61) and Prakken (2016).

machine learning (Hertz et al. 1991: 94-7; Minsky and Papert 1969). To see what the problem is, try to draw a straight line separating the black and white dots:

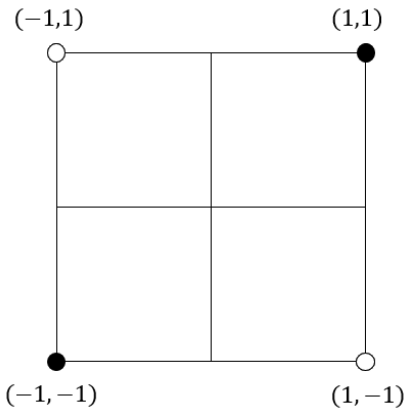

Figure 4: The Exclusive Disjunction

That it is impossible to draw a line separating the 'ought' and 'ought not' cases is equivalent to the claim that there exist no weights $\mathbf{w}$ and bias b such that $\Sigma w_i x_i + b > 0$ for $\mathbf{x} = (-1,1)$ and $\mathbf{x} = (1,-1)$; and $\Sigma w_i x_i + b \leq 0$ for $\mathbf{x} = (1,1)$ and $\mathbf{x} = (-1,-1)$. So, the additive model cannot handle cases in which what I ought to do depends on an exclusive disjunction of two factors. For similar reasons, the additive model cannot handle cases in which what I ought to do depends on a biconditional of two factors, i.e. I ought to φ just in case both or neither of two factors obtain. The remaining 14 binary Boolean functions are linearly separable, so the model is capable of representing reasons-interactions that correspond to these functions.

Dancy and friends would have a field-day if they realised the counterexample explosion that arises once more factors are added (c.f. Crama and Hammer 2011: 412-3). In three-factor cases, there are 256 Boolean functions and 152 of them are non-separable. Thus about 60% of the logically possible ways in which reasons might interact are ruled out. In four-factor cases, there are 65536 Boolean functions, and 63654 are non-separable. Thus about 97% of logically possible reasons-interactions are ruled out. At the risk of overegging the pudding, in five-factor cases there are 4294967296 Boolean functions, and 4294872724 non-separable functions. Thus about 99.99% of possible reasons-interactions are ruled out in cases involving five different factors. This trend is alarming. As the number of morally relevant features increases, the additive model rules out the vast majority of ways in which reasons might admissibly interact to determine right-action.

Once the counterexamples to the additive model in two-factor cases are situated in their proper context, these cases seem less troubling. These cases are problematic only if we accept that *any* Boolean function can in principle represent an instance of reasons aggregation. But why should we accept this claim? The fact that we are using Boolean functions to model *aggregation* suggests that only a small subset of these functions that have certain desirable properties will be relevant. Take voting. This is another context in which Boolean functions are used to model aggregation. What is being aggregated *here* are people's preferences as opposed to moral reasons. The Boolean functions that are applicable in voting theory are those that satisfy desirable properties such as weak monotonicity. Here what it means for a function to be weakly monotonic is that if one candidate is among the winners in the election, then that candidate does not become a loser if one person switches their vote in that candidate's favour (Mackie 2003: 55-6). The fact is that additively representable Boolean functions have properties that are well-suited to aggregation – be it aggregation of preferences, reasons, or the processing needs of different parts of a system in distributive computing (Crama and Hammer 2011: 405-8). Thus I do not think that we should abandon the additive model because of these alleged problem cases. There are compelling structural reasons to stick with additive functions.

I realise that this is not a decisive argument for the additive model. But I think it is sufficient for us to shelve Kagan's point about the strong separability of reasons. Kagan's other point was that trolley cases presuppose that the governing function is ubiquitous, i.e. if a feature is morally relevant in one case, then it is also morally relevant in all the cases where it is instantiated. There are two plausible responses here. One is to follow Roger Crisp (2003) and argue that the model applies only to ultimate reasons such as those grounded in *justice* or *harm*. The cost to this response is that often people wish to use trolley cases to discuss more specific features such as the killing and letting-die distinction. One stronger response is to point out that the trolley case method does not require *ubiquity*, but instead that the relevant cases and the intended target cases in the real-world are sufficiently similar. For instance, trolley cases might be relevant to the ethics of killing in AV collisions, but at the same time may not be relevant to killing in more remote cases such as war. What matters, then, is that we have no reason to believe that the moral considerations in the target systems are categorically different from those considered in the trolley case. The assumption that the governing function is ubiquitous is not required.

2.1.3. Using Trolley Cases

Given what I have said so far, it should be clear that AVs need not encounter trolley cases in practice for them to be relevant to what morality requires in AV decisions. I shall now say something about how trolley cases might be used to inform the ethics of AVs in practice. Recently, de Freitas, Anthony and Alvarez claimed:

> The main safety goal for any driver – human or machine – is to avoid harm [...] Unfortunately, both humans and today's best computer systems are imperfect at it. Even so, the substantial improvements that we rightfully expect from future AV systems are utterly unlikely to come from considering trolley dilemmas (de Freitas, Anthony and Alvarez 2019: 4).

This is an empty criticism. Trolley cases are not a programming tool, so it is obvious that trolley cases will have no useful application in this domain. But trolley cases are relevant to the claim that avoiding harm ought to be the main safety goal for AVs. This claim, interpreted in the terminology of the model sketched above, holds that the only property which makes a difference to the moral permissibility of the AV's acts is the property of causing harm. Trolley cases provide good reason to believe that this claim is false. Consider Philippa Foot's (1967/2002) original:

> *Trolley Driver:* A trolley's brakes fail. The driver can continue on the same track and kill five workmen; or steer the trolley onto another track, saving the five workmen but killing one workman on the other track.

And

> *Transplant:* A surgeon can kill one patient and use their organs to save the lives of five other patients.

If de Freitas, Anthony and Alvarez are correct that harm is all that matters, then the trolley driver is morally required to kill one person to save five; and the surgeon is morally required to kill one patient and harvest their organs to save five others. In both cases, these actions would cause the least harm. But whilst most people have the intuition that it is permissible to redirect the trolley in Trolley Driver, most people also have the intuition that it is morally impermissible to kill one to save five in Transplant. It would be a serious moral error if the surgeon killed one patient with the express intention of harvesting their organs to save five others.

Presumably, what explains the difference in our intuitions is that the cases differ in some morally significant respect. In other words, there is a property other than

harm contributing to the moral permissibility of the actions. But if this is true, then de Freitas, Anthony and Alvarez are incorrect that harm is the *only* property of acts which makes a difference to moral permissibility. So, one application of trolley cases to the general moral design problem is that trolley cases can be used to argue against theories about which properties make a difference to moral permissibility.

It might be objected that Trolley Driver and Transplant are too divorced from the context of AV collisions to make inferences from these cases to AV collisions. But it is possible to make analogous cases which are closer to the context of AV collisions. For example, suppose that a motorcyclist is skidding across the road towards a crowd of pedestrians on the pavement. The AV can brake, in which case the motorcyclist will skid into the pedestrians and cause their deaths. The AV could also accelerate into the motorcyclist, in which case the motorcyclist would be killed, but the skid would be deflected and the pedestrians would be unharmed. Presumably, it is morally permissible for the AV to brake here. It is too demanding to suppose that the AV is morally required to intervene and kill the motorcyclist. But if de Freitas, Anthony and Alvarez are right that harm is all that matters, then the AV is morally required to kill the motorcyclist. The salient point here is that properties other than harm seem to make a difference to the moral permissibility of the AV's acts in ideal cases. It seems that whether or not the harm is *done* or *merely allowed to happen* is morally relevant. Presumably, the properties of doing harm and allowing harm to happen are instantiated by some acts available to AVs in the real-world. So, absent good reason to believe that the relation between these properties and moral permissibility is *radically different* in real-world cases, the view that avoiding harm is all that matters in AV collisions ought to be discarded.

In making precise the features of acts which affect their moral permissibility in cases like Trolley Driver and Transplant, we can also formulate positive arguments about which features of acts AVs should take into account when making decisions. Foot's explanation of the conflicting intuitions in Trolley Driver and Transplant appeals to *positive* and *negative* duties. Foot argues that Trolley Driver involves a conflict between two negative duties: the duty not to kill one and the duty not to kill five. The duty not to kill five is stronger than the duty not to kill one, so it is morally permissible to kill one to save five others. In contrast, the conflict in Transplant is between the negative duty not to kill one and the positive duty to aid five others. It is better, on Foot's view, to let five people die than it is to kill one person (Kamm 2016: 15-6; Thomson 2008: 360, 2016: 114). So, Foot's explanation of why it is permissible to kill one to save five in Trolley Driver but not in Transplant is that Trolley Driver is a choice between *killing one* and *killing many*,

whereas Transplant is a choice between *killing one* and *letting many die*. If Foot's account of the difference in our judgements across Transplant and Trolley Driver is correct, then plausibly the same point holds in the context of AV decisions.

To conclude: The Not Going to Happen Argument holds that trolley cases are irrelevant to the general moral design problem *because* AVs will not encounter trolley cases in the real world. This argument falsely assumes that trolley cases are relevant only if AVs will encounter these cases in practice. What matters for the relevance of these cases to practical ethical dilemmas is that the acts in real-world AV driving instantiate some of the morally relevant properties which trolley cases are concerned with. It is not unreasonable to suppose, for example, that whether an AV kills someone or merely allows them to die makes a difference to the moral permissibility of the AV's actions. So, trolley cases may be relevant to the moral design problem *even if* these cases will not arise in practice. Hence the Not Going to Happen Argument against the relevance of trolley cases is unsound.

2.2. The Moral Difference Argument

According to the *Moral Difference Argument*, trolley cases and real-world collisions differ in some morally significant respects; and these differences render trolley cases less helpful than we might hope when tackling the general moral design problem.

One version of this argument holds that trolley cases ignore certain features of the AV's acts that make a difference to the rightness of the AV's acts in real-world collisions (Nyholm and Smids 2016: 1282–84). For example, trolley cases abstract away information about who is morally responsible for the collision; and about the special ties that may or may not exist between the affected parties in the collision.

This argument presents an important criticism for some uses of trolley cases in the ethics of AVs. Derek Leben's (2017) 'Rawlsian' algorithm is a good example. Leben's algorithm is a set of instructions for how AVs should behave in collisions. The trouble is that Leben takes trolley cases as *models* for collisions. The algorithm takes as its input information about the survival probabilities of each affected party conditional on each act available to the AV. It then outputs what the AV morally ought to do on the basis of this information alone. Accordingly, Leben's algorithm is insensitive to facts that have prima facie moral relevance. This is a serious defect in Leben's algorithm. But we should not throw the baby out with the bathwater.

First, I do not think that trolley cases are the correct object of criticism. It is not trolley cases that overlook morally relevant considerations. That Leben uses trolley cases as inputs to his algorithm is his prerogative. This is not a failing on the part of the trolley case method. It is a failing on Leben's part, and on the part of anyone else using trolley cases as inputs to collision algorithms. Second, as I explained in the last section, the trolley case method aims to hold fixed all features of a situation bar one, to see if permuting that feature makes a difference to the moral status of the AV's acts. It is one thing to *forget about* morally relevant considerations like who is responsible for the collision and who has special ties to whom. It is another thing to stipulate that there are no special ties, or that no one is responsible, so that we can focus our attention on whether or not particular facts have moral relevance. That is what happens in trolley cases when they are used in the right way. So, I am unconvinced that this first version of the argument is a problem for trolley cases.

There is a better version of the moral difference argument. This argument holds that the moral considerations in trolley cases are categorically different to the moral considerations in real-world collisions. So, trolley cases are less helpful than might be expected for addressing the general moral design problem. This argument has two steps. The first is to establish a non-normative difference between trolley cases and real-world collisions. Standardly, it is argued that in trolley cases the AV knows the outcome of the collision conditional on each act. But in real-world cases, the AV at best has a probability distribution over each outcome conditional on each act. In short, AV collisions involve risk, and perhaps other forms of uncertainty (Himmelreich 2018: 676-7; Nyholm and Smids 2016: 1286). The second step holds that the presence of risk in collisions renders the moral dilemmas in these cases different in kind to those in trolley cases. Sven Nyholm and Jilles Smids write:

> Reasoning about risks and uncertainty is categorically different from reasoning about known facts and certain outcomes. The key concepts used differ drastically in what inferences they warrant. And what we pick out using these concepts are things within different metaphysical categories, with different modal status (e.g. risks of harm, on one side, versus actual harms, on the other) (Nyholm and Smids 2016: 1286).

I agree that there is a difference between reasoning about risks and reasoning about certain outcomes. However, I do not think that this difference is sufficient to challenge the relevance of trolley cases to the general moral design problem.

Consider Leonard Savage's (1972) model of decision-making under risk. There are several ingredients to put in place. First, there is a set of mutually-exclusive

and exhaustive *states*. Here a state is 'a description of the world, leaving no relevant aspect undescribed' (Savage 1972: 8). I denote these states $s_1, s_2, \ldots, s_n \in \mathbf{S}$. One of these states is the actual state of the world. The agent does not know which. Second, a *consequence* is something that might happen to the agent or be of significance to the agent. I denote the possible consequences $c_1, c_2, \ldots, c_m \in \mathbf{C}$. Savage (1972: 14-15) understands the *acts* that the agent must choose between as functions from $\mathbf{S}$ into $\mathbf{C}$. The intuition here is that what matters for characterising an act is what the consequences of the act will be under each possible state. For Savage, two acts that have the same consequences under each state might as well be considered the same act. I denote the acts $f_1, f_2, \ldots, f_K \in \mathbf{F}$, where each f_k is a function $f_k: \mathbf{S} \rightarrow \mathbf{C}$.

The agent has a preference relation $\preccurlyeq$ over the acts in $\mathbf{F}$. Here $f \preccurlyeq g$ if, and only if, act f is at most as good as act g from the agent's point of view. Savage proved that if $\preccurlyeq$ meets certain conditions, three claims are true.[14] First, the agent's degrees of belief in the propositions that each state s is the actual state of the world can be represented with a unique probability function.[15] Second, $\preccurlyeq$ can be represented with a family of utility functions that are unique up to positive linear transformations.[16] Third, the probability function and the utility function can be used to make an expected utility function such that $f \preccurlyeq g \leftrightarrow EU(f) \leq EU(g)$. The expected utility of an act is the weighted sum of utilities conditional on the act being performed under different states. The weights are given by the probability of each state.

$$EU(f) = \sum_{i=1}^{n} P(s_i)u\big(f(s_i)\big)$$

The function EU over acts is well-defined only if the utility function u over consequences is well-defined. Thus I am unconvinced by Nyholm and Smids' claim that 'reasoning about risks … is categorically different from reasoning about

[14] The axioms are: (1) That $\preccurlyeq$ is complete and transitive. (2) The sure thing principle, which roughly says that each (set of) outcome(s) of an act is evaluable independently of other (sets of) outcome(s) of that act. (3) That the agent is certain that an event (set of states) is not the case if, and only if, conditional on that event the agent $f \preccurlyeq g$ and $g \preccurlyeq f$ for all $f, g \in \mathbf{F}$. (4) That knowing which state is the case does not impact the agent's preferences over outcomes. (5) Suppose that f produces consequence X under event E, and X' under $\neg E$; and that f' has consequences X and X' under events F and F'. Further suppose that g produces Y and Y' under E and $\neg E$; and Y and Y' under F and $\neg F$. Then if $X \preccurlyeq X'$, $Y' \preccurlyeq Y$, and $f \preccurlyeq f$, then $g' \preccurlyeq g$. (6) There exists at least one pair of acts $f, g \in \mathbf{F}$: $f \prec g$ (Savage 1972: esp. 17-26).

[15] The bearers of probability are *events*, which are elements of $\mathcal{P}(\mathbf{S})$, i.e. the set of subsets of $\mathbf{S}$. Hence $P: \mathcal{P}(\mathbf{S}) \rightarrow [0,1]$. The reason that the probability function is defined over events is that the domain of the probability function needs to be closed under the standard logical operations, i.e. be an algebra, to satisfy the probability axioms (c.f. Savage 1972: 10-13).

[16] This means the utilities are unique up to multiplying by a constant and adding a scaler.

known facts and certain outcomes.' In Savage's theory, there is no point talking about *expected utility* unless we have a clear idea of *utility*. The former depends on the latter. The aim of trolley cases is to discern which features of a case are morally relevant, i.e. to discern the properties that the moral betterness relation $\preccurlyeq_M$, and the moral utility function, u_M, that represents it, are sensitive to. Accordingly, I do not see how we can approach the morality of risky decisions without having a plausible idea of what morality requires in non-risky cases (c.f. Crisp 2006: 39-40). Expected moral utility is nonsensical in the absence of well-defined moral utilities.

There are various ways to push back against this argument. First, it is not clear that the morality of decisions under risk is best understood in terms of maximising a moral expected utility function. The presence of risk in a choice might itself have moral relevance. This would be true if facts about the variance of moral utilities under the possible outcomes of an act made a difference to the act's moral status. However, this kind of response is dialectically inert. This is because risk-weighted decision theories still pitch the agent's evaluation of risky prospects as a function of the agent's utilities under different states of the world and the probabilities of those states (c.f. Buchak 2013: 48-59). Hence the agent's risk-weighted expected utility function is well-defined only if their utility function is well-defined. Second, certain moral considerations arise in risky cases that do not arise in non-risky cases. For example, the circumstances in which an agent is morally permitted to dismiss a risk as morally insignificant (c.f. Bjorndahl et al. 2017). This kind of response does not work either. The argument I am presenting does not rely on the claim that the moral considerations in risky cases are identical to those in non-risky cases. What I am claiming is that claims about the morality of risky prospects are nonsensical in the absence of an underlying axiological commitment that *something* matters; and considering hypothetical non-risky cases may inform our axiological commitments.

2.3. The Impossible Deliberation Argument

According to what I call the *Impossible Deliberation Argument*, AVs cannot deliberate in accordance with answers to the general moral design problem based on trolley cases. This is because solutions to trolley cases presuppose a 'top-down' approach to AV engineering (Gurney 2015: 208; Himmelreich 2018: 675; Nyholm 2018b: 5). What it means for AVs to be programmed 'top-down' is roughly that in collisions the AV follows a set of rules that are expressible in first-order logic (c.f. Alaieri and Vellino 2016: 161-2; Allen et al. 2005: 149-51). The foil for the 'top-down' method

is the 'bottom-up' method. Here the AV's decisions are governed by a connectionist algorithm that has no explicit rules. What is problematic is that AVs are standardly programmed in accordance with the 'bottom-up' approach. Hence it is unclear how answers to the moral design problem based on trolley cases can be implemented.

Himmelreich puts the point as follows:

> [trolley] cases naturally lend themselves to a top-down design approach. But given the current prominence of the bottom-up approach in artificial intelligence in the form of neural networks, there is a risk of a discontinuity of approaches between ethics and engineering (Himmelreich 2018: 675).

This is not a very good argument. I have two responses. First, consider traffic laws. Presumably, AV designers know how to design AVs such that AV behaviour is consistent with traffic laws. It would be problematic if this were false. However, traffic laws naturally lend themselves to a 'top-down' approach to AV engineering. The rules of the road are in effect a big list of conditional statements. If the speed limit is such-and-such, then the vehicle's speed should be at most such-and-such. If the pedestrian is waiting at the crossing, then stop to let the pedestrian cross. Clearly, if solutions to the general moral design problem based on trolley cases lend themselves to a 'top-down' approach to AV design, then implementing these solutions into AVs will be at most as difficult as implementing road-traffic laws.

Second, I take it that the principal role for philosophers in AV ethics is to offer plausible criteria for rightness for AVs. It is not to provide decision procedures. Here a criterion of rightness is a principle or set of principles that explain the moral status of the AV's actions; and a decision-procedure is an algorithm for deliberation (Brink 1986: 421; c.f. Bales 1971; Crisp 1992). Trolley cases are relevant to criteria for rightness. The point of using trolley cases is to decide between competing accounts of the right-making features of the AV's acts. These cases are intended to determine *what matters morally* in AV decisions, be it utility, fairness, justice, rights, responsibility, justifiability to the affected parties, or some combination of these. I do not see how the use of connectionist algorithms in AVs impacts our discussions about the right-making properties of the AV's actions in collisions. I grant that insofar as philosophers are interested in designing or informing AV designers about the ethical considerations relevant to candidate AV decision-procedures, then it is important for philosophers to take into account the nature of AV decision systems. This is what I do in Chapter 6. But for the more basic question of what matters in collisions, it is unclear how the use of neural networks for AVs presents a problem.

2.4. The Wrong Question Argument

I now turn to what I take to be the most interesting criticism of trolley cases and their application to the ethics of AVs. According to the *Wrong Question Argument*, trolley cases may well be relevant to what morality requires in AV driving. But the task of finding principles for the regulation of AV behaviour is not a moral problem. It is a political problem. So, trolley cases are inappropriate for theorising about principles for the general regulation of AV conduct (Himmelreich 2018: 675-6).

This is a deeply interesting challenge. First, Himmelreich (2018: 676) argues that there is no unanimous agreement about which moral principles are true. So, any answer to the general moral design problem based on trolley cases is unlikely to receive 'broad societal acceptance' (c.f. Santoni de Sio 2017; Keeling 2018a). Himmelreich thinks that broad societal acceptance is a necessary condition for a successful solution to the general moral design problem. So, what is required is not an answer to what morality requires in AV driving, but instead an answer to a social choice problem. In his words: 'A trolley case prompts us to make an individual choice when what we in fact face is a social choice. What seems needed is a kind of compromise to overcome disagreements over issues of value' (2018: 676).

I want to flag a couple of things about this argument. First, Himmelreich thinks that 'what counts as the right choice [for AVs in] dilemma situations is essentially contested' (2018: 676). Thus as Himmelreich sees it, AVs present something of a predicament. We must find broadly agreeable principles for the general regulation of AV behaviour, where these principles are justifiable to the public despite their reasonable disagreements about how AVs morally ought to behave (c.f. Rawls 2005). Second, Himmelreich thinks that an approach to AV ethics that is based on considerations about trolley cases is somehow at odds with this political conception of how we ought to go about finding principles for AV behaviour. What I think is Himmelreich's concern is that a characteristically moral approach to AV ethics risks imposing moral values on people where those values are not universally endorsed.

To fully get to grips with this argument, I shall make explicit the political philosophy that underpins it. The argument is based on John Rawls' (2005) theory of political liberalism. Rawls took as basic the fact that people have different views about morality, religion, and so on. That people have different moral outlooks is a problem for liberal states because liberal states must legislate on moral questions. The task for liberal political theory is to account for the legitimacy of these laws, i.e. to explain why it is legitimate for the state to coerce people who share different

moral beliefs to adhere to legislative solutions to moral questions. Rawls' solution to this problem is that it is legitimate for states to exercise political power only if that power is exercised in a way that reasonable citizens can endorse. Consider,

> citizens are reasonable when, viewing one another as free and equal in a system of social cooperation over generations, they are prepared to offer one another fair terms of cooperation according to what they consider the most reasonable conception of political justice; and when they agree to act on those terms, even at the cost of their own interests in particular situations, provided that other citizens also accept those terms (Rawls 2005: 446).

Rawls held that in a democratic culture comprising reasonable citizens, public officials should strive to provide public justifications for constitutional, judicial, and legislative decisions. These justifications appeal to political values such as freedom and equality, that arise from an overlapping consensus in the divergent moral outlooks of the citizens. Thus Rawls claims that 'our exercise of political power is proper only when we sincerely believe that the reasons we would offer for our political actions … are sufficient, and we also reasonably think that other citizens might reasonably accept those reasons' (2005: 446–7). Rawlsian public justification constitutes a pragmatic and deeply interesting approach to the ethics of AVs.

With Himmelreich's foundational commitments on the table, we are now in a position to advance some criticisms. But before I do, I should like to register what I have learned from Himmelreich's argument. The first time that I came across this argument, I thought that it relied on a naïve picture of moral philosophy on which the answers to applied ethical disputes are formulated by the flat-footed application of first-order normative ethical theories to particular domains. This was a serious misunderstanding. I now appreciate that Himmelreich is correct to put pressure on moral philosophy as a method for addressing questions of great social importance. There is at times a failure on the part of philosophers to appreciate that the question of which principles should regulate AV behaviour at all times on the road is not simply a theoretical question. The arguments that we advance need to be the sorts of arguments that can shape legislation and policymaking. This requires a degree of methodological compromise; and a willingness to understand the differences between our approach to these questions and that of policymakers. The suggestion that moral philosophy might benefit from taking account of the fact that reasonable disagreement exists about difficult moral questions is an important proposal.

I shall now advance some criticism. What I want to suggest is that Himmelreich is broadly correct about the political approach. But when the implications of this

approach are made precise, it is clear that trolley cases are not the natural foil. First, I shall make clear what Rawls took as the foil to his account of political justification. Rawls takes political justification to stand in contrast to dogmatism. Consider,

> Since [...] public reason specifies at the deepest level the basic political values and specifies how the political relation is to be understood, those who believe that fundamental political questions should be decided by what they regard as the best reasons according to their own idea of the whole truth – including their religious or secular comprehensive doctrine – and not by reasons that might be shared by all citizens as free and equal, will of course reject the idea of public reason. Political liberalism views this insistence on the whole truth in politics as incompatible with democratic citizenship and [...] legitimate law (2005: 447).

What Rawlsian political justification stands in contrast to is political officials imposing their conception of the good on other people in a way that is not justifiable to them on grounds that they can reasonably accept as citizens in a democracy. Thus the burden on Himmelreich is to explain what it is about trolley cases that renders these cases symptomatic of the political dogmatism that Rawls disavows. This is a difficult task. I cannot see that there is anything *essential* to trolley cases that requires them to be used in conjunction with dogmatic justifications for views about the principles for governing AV behaviour. Presumably, trolley cases can be used alongside dogmatic or non-dogmatic justifications for AV ethics legislation.

In response, Himmelreich might claim that on political questions about which there is much disagreement there is something inherently dogmatic about putting forward moral arguments in favour of certain legislative solutions. This strikes me as a mistake in two respects. First, the Rawlsian view does not preclude the social conversation about legislative matters including opinions grounded in conceptions of the good that are not shared among all citizens. Indeed, Rawls claims that 'reasonable comprehensive doctrines, religious or non-religious, may be introduced in public political discussions at any time, provided that in due course proper political reasons [...] are presented that are sufficient to support whatever the comprehensive doctrines introduced are said to support' (Rawls 2005: 462). Second, the method of case-judgements that Himmelreich takes as the foil for his political approach is not in itself a moral doctrine. It is a method. There is no conception of the good that is common to all users of trolley cases. These cases are designed to elicit shared intuitions about the plausibility of different moral theories. Thus I do not see how trolley cases represent the natural foil to Himmelreich's political view.

There are, I think, two respects in which this response to the Wrong Question Argument can be strengthened. The first is to reflect on what the political approach to AV ethics might amount to in practice. The most extreme form of this view is to pitch the problem of finding principles for the general regulation of AV behaviour as a social choice problem. Social choice theorists are interested in '[whether] it is formally possible to construct a procedure for passing from a set of known individual tastes to a pattern of social decision-making, the procedure in question being required to satisfy certain natural conditions' (Arrow 1963: 2). The standard way of making this question precise requires us to suppose that each individual in society has a preference ordering over a set of alternative policies. The challenge is then to describe a *social welfare function*, which takes the individual preference orderings as an input, and outputs a collective preference ordering over the policies. Social choice theorists dispute which formal properties the social welfare function should satisfy (*Ibid.*: 23; Sen 1970: 35-6). If the general moral design problem is construed as a social choice problem, then the values of all the individuals in society are taken to *uniquely determine* the principles for regulating AV behaviour.

On this extreme form of Himmelreich's political approach, there is good reason to think that trolley cases, and moral philosophy in general, have an important role to play in the realisation of the political approach. This is because on any plausible theory of deliberative democracy, it is considered a virtue of the democratic process that the electorate is well-informed (Brennan 2011). I take it that familiarity with the different moral theories, and what these theories imply about particular cases, is a necessary condition for being *informed* about the general moral design problem. Presumably, a social choice approach to the problem that *excluded* the use of trolley cases and other resources from moral theory would be greatly impoverished. Hence even the most extreme form of the political approach leaves room for trolley cases. There also obvious respects in which this point holds for less extreme forms of the political view that involve different degrees of public participation in the legislative process. For example, citizens' assemblies, focus groups, public debates, and so on.

The second point: I suspect that Himmelreich overplays the extent of moral disagreement that exists in practice. I take it that the general moral design problem raises questions about which there exists a sufficiently broad consensus to invoke moral arguments for and against candidate solutions. Himmelreich is sceptical of legal moralism. But for Rawls, legal moralism is problematic only insofar as it targets basic political equality or the liberty of citizens on grounds that the targets of those arguments do not accept. For example, a Rawlsian may object to laws against gay marriage or free religious worship based on dogmatic moral arguments.

Rawlsian political liberalism does not exclude invoking moral arguments to support laws on murder and self-defence. There is a lot of consensus on moral permissions in these areas, even if people disagree about the details. I think that AV ethics is similar. Moral arguments – involving trolley cases or not – have an important role to play in the public conversation on the regulation of AV behaviour.

2.5. Conclusion

I have made the case for the relevance of trolley cases to AV ethics. I am hesitant to attach too much significance to methodological arguments. The best test of trolley cases is their use in practical moral theory building. I hope that my use of these cases in the following chapters illustrates the usefulness of these cases.

3. Legal Necessity and Pareto Efficiency[17]

3.1. Introduction

Suppose that an automated vehicle (AV) encounters a scenario where harming at least one person is unavoidable and a choice is required about how to distribute harms between multiple persons whose interests are in conflict. What does morality require here? How should the AV be programmed to distribute benefits and burdens between the affected parties? I call this the *moral design problem.*

Filippo Santoni de Sio (2017) defends a *legal-philosophical approach* to the moral design problem. According to Santoni de Sio, we ought to take the criminal law as a starting point for our moral theorising. The rationale behind this approach is that answers to the moral design problem are based on moral principles. But people disagree about which moral principles provide the correct account of justified harm; and this disagreement is likely to prevent us reaching a social consensus on the solution to the moral design problem. Santoni de Sio believes that the criminal law can resolve this disagreement. He formulates an answer to the moral design problem based on the legal doctrine of necessity as it features in Anglo-American criminal law. Santoni de Sio's answer is in effect a form of rights-based deontology.

In this chapter, I argue that we should reject the legal-philosophical approach. First, I argue that Santoni de Sio's answer to the moral design problem does not achieve the aim of the legal-philosophical approach. This is because his answer takes for granted the truth of certain moral principles which, at least, utilitarians have reason to reject. Essentially the problem is that Santoni de Sio postulates a rival moral theory based on the criminal law and provides no reason for proponents

[17] This chapter is based on my paper 'Legal Necessity, Pareto Efficiency and Justified Killing in Autonomous Vehicle Collisions' in *Ethical Theory and Moral Practice*, 2018, 21(2), pp. 413-427. I have kept most of the original paper intact. But whilst the original paper argued in favour of the Restricted Pareto Principle, I now argue against it.

of other moral theories to accept his view. Hence Santoni de Sio's account of permissible harm does not resolve our moral disagreements; and there is no obvious reason to think that it can bring us to a consensus on the moral design problem.

Second, I argue that the legal-philosophical approach is salvageable to a limited extent. I offer an alternative reading of the legal doctrine of necessity and what it implies about unavoidable collision scenarios. My account appeals to

> *The Restricted Pareto Principle:* The AV is justified in killing if the outcome on which the person is killed is a unique Pareto efficient outcome.

Here a Pareto efficient outcome is an outcome such that there exists no other outcome on which all parties are at least as well-off and at least one party is strictly better-off. An outcome is uniquely Pareto efficient if it is the only Pareto efficient outcome in the set of possible outcomes. I argue that utilitarians, contractualists and deontologists can accept the Restricted Pareto Principle as a partial answer to the moral design problem, even if they disagree about which moral principles are true. This is because the Restricted Pareto Principle stipulates a *condition under which* an act of killing is justified as opposed to *that in virtue of which* the act is justified. Santoni de Sio's mistake was to postulate a rival moral theory. What I do is present a condition that all three moral theories can assent to despite their disagreements about which moral principles explain why this condition is true.

Third, I argue that the Restricted Pareto Principle is at best a principle which any plausible answer to the moral design problem should imply. It is not a complete answer to the problem because it applies in very few cases. In short, the problem is that agreement between the moral theories comes at the cost of *decisiveness*, i.e. the best moral theories disagree about many cases, hence a view based on the conditions under which these theories agree will be silent about what ought to be done in many cases. Last, as the Restricted Pareto Principle is the most plausible consequence of the legal-philosophical approach to the moral design problem, I conclude that we ought to reject the legal-philosophical approach in favour of some other approach.

Here is the plan. In §3.2, I articulate the legal-philosophical approach. Then I distinguish two readings of the approach. On the *wide reading*, the aim is to use the law to resolve our disagreements about which moral principles offer the correct account of justified harm. On the *narrow reading*, the aim is to use the law to bring us to a consensus on the moral design problem despite our disagreement about which moral principles offer the correct account of justified harm. In §3.3, I explain Santoni de Sio's answer to the moral design problem, which is based on the legal doctrine of necessity. I then argue that, on either reading of the legal-philosophical

approach, Santoni de Sio's answer falls short of its aim. In §3.4 and §3.5, I offer the Restricted Pareto Principle as an alternative reading of the doctrine of necessity; and I argue that three of our best moral theories can agree on this principle as a partial solution to the moral design problem. Hence the Restricted Pareto Principle provides a partial vindication of the narrow reading of the legal-philosophical approach. In §3.6, I argue that we ought to reject the Restricted Pareto Principle. Because the Restricted Pareto Principle is the most plausible consequence of the legal-philosophical approach, we should reject this approach.

3.2. The Legal-Philosophical Approach

Santoni de Sio's (2017) legal-philosophical approach aims to overcome what we can call the *problem of moral disagreement*. I first explain the problem. I then distinguish two readings of how the legal-philosophical approach aims to solve this problem.

Answers to the moral design problem are based on moral principles, in the sense that each answer invokes at least one moral principle to explain the moral status of the AV's acts. For example, Bonnefon et al. (2016) present a utilitarian answer, according to which the AV should be programmed to minimise loss-of-life in collisions. The explanatory principle here is some form of the Principle of Utility, i.e. acts are right just in case they maximise utility impartially considered.[18] Leben (2017) defends a contractualist view, according to which AVs should act so as to maximise the survival probability of the worst-off party. The explanatory principle here holds that the AV should behave in accordance with principles that the parties in the collision would rationally consent to under fair bargaining conditions.

People disagree about which moral principles are true. Santoni de Sio thinks that our disagreements about morality will prevent us from reaching a consensus on the moral design problem. In his words, 'both lay people and philosophers disagree about what is morally prohibited, permissible or obligatory in scenarios where fundamental interests and values are at stake, so that neither experimental ethics nor philosophical ethics seem at the moment able to offer [an uncontested solution to the moral design problem]' (Santoni de Sio 2017: 412). Presumably, Santoni de Sio believes that a solution to the moral design problem is acceptable only if all or almost all people accept that solution (c.f. Bonnefon et al. 2016: 1573;

[18] I should flag that Bonnefon et al. (2016) present a particular interpretation of what the principle of utility implies about the morality of AV decision-making. But this is a rather crude interpretation. See the 'Utilitarianism' section of Chapter 1 for further discussion.

Himmelreich 2018: 675-676). If *this* is true, then we can solve the moral design problem only if we provide a solution that receives near universal acceptance.[19]

I shall make some remarks to tighten-up the problem.[20] First, Santoni de Sio's problem of moral disagreement should not be mistaken for the problem of moral uncertainty (c.f. Ord and MacAskill 2018; MacAskill 2016; Oddie 1994; Lockhart 2000; Ross 2006; Guerrero 2007; Sepielli 2009; Weatherson 2014). What theories of moral uncertainty describe is how we ought to act when uncertain about which first-order normative ethical principles are true *given that* we have a general aim of behaving in accordance with the requirements of morality. Hence the sort of views at issue in moral uncertainty explain how to compromise between the prescriptions of different moral theories taking into account our credences in each theory. The problem for Santoni de Sio is that we need 'reasonable guidelines for solving the ethical issue of the regulation of the programming of AVs' (2017: 413); but that our moral disagreements prohibit moral philosophy from providing those guidelines. What is required from Santoni de Sio's point of view is a method for resolving moral disagreements in the face of practical challenges that require definitive answers. The aim is not so much one of minimising the chances of wrongdoing, as is the case with moral uncertainty. Instead the aim is something like that of reaching a defensible compromise between the views that different people hold about morality.

That concludes the challenge. We can ask: How does the legal-philosophical approach aim to overcome this methodological challenge? According to the *wide reading*, the law can bring us to a consensus about which moral principles provide the correct account of justified harm. The idea behind the wide reading is that the law contains novel insights that moral philosophers could take into account, such that taking into account these insights could help us to overcome our disagreements about morality. In support of this reading, Santoni de Sio claims:

> The main methodological idea behind this approach is John L. Austin's (1956) suggestion that legal reasoning may be a sharp instrument of clarification on complicated philosophical questions. According to Austin, the reflections of

[19] This approach might well seem unreasonably strong. It is not standard practice to find policies that all people in fact agree to. We do not think that taxation policies are a problem because some people are libertarians, for example. I think this criticism is correct. But to make a convincing case for this criticism it is necessary to show that the legal-philosophical approach fails even on its best formulation. This is what I hope to show in what follows.

[20] In addition to these points, it is worth noting that there is a similarity between Santoni de Sio's (2017) approach to the moral design problem, and to Derek Parfit's (2011a) project in *On What Matters*. Parfit sought to show that, correctly understood, three of our best moral theories are 'climbing the same mountain on different sides.' Though Santoni de Sio's execution of the legal-philosophical approach differs in several important respects to Parfit's approach, my development of Santoni de Sio's view is similar in spirit to Parfit's approach.

> lawyers – with their standing attention to real-life cases [...] may offer a fresh start to address difficult philosophical problems. Whereas I do believe that looking for fresh solutions to new or hard ethical problems is ultimately a philosophical enterprise [...] I also think that philosophical reflection may sometimes benefit from considering legal principles and norms (2017: 413).

There is a second reading of the approach. According to the *narrow reading*, the law can bring us to a consensus on the moral design problem despite our disagreements about which moral principles give the correct account of justified harm. Here, the aim is not for the law to solve our moral disagreements. The aim is instead to use the law to bring us to a consensus on the moral design problem in the absence of any general agreement about which moral principles offer the correct account of justified harm. In support of this reading, Santoni de Sio writes:

> [...] legal norms are often the result of a combination of abstract moral principles and practical considerations deriving from the close observation and comparative analysis of real cases; moreover, legal norms are often an explicit attempt to cope with the fact of disagreement about general normative principles by finding a "reasonable compromise between principles and interests in contrast" (Hart 1961: 128) (*Ibid.*).

It is unclear to me which of these readings is Santoni de Sio's intended reading. Because Santoni de Sio is ambiguous between the two, I shall argue that his solution to the moral design problem fails on *either* of these readings.

3.3. The Doctrine of Necessity

In this section, I first outline the doctrine of necessity. I then explain how Santoni de Sio uses this doctrine to formulate an answer to the moral design problem. Finally, I argue that on both the narrow and the wide readings of the legal-philosophical approach, Santoni de Sio's answer to the moral design problem fails to achieve the aim of the approach.

3.3.1. The Doctrine of Necessity

What is the doctrine of necessity? In broad terms, necessity is a legal defence. It stipulates sufficient conditions for a defendant to be absolved of criminal liability. Necessity arises in situations where the defendant faces

> [a] choice between values, protected interests, etc., where one of the defining features is that the prospective defendant is free to choose which course to take (Bohlander 2006: 150-1).

The necessity defence recognises that, sometimes, a person will pursue a course of action which is ordinarily regarded as criminal, but their action should not be regarded as criminal because the decision they faced made it *necessary* for them to violate the ordinary wording of the law (Arnolds and Garland 1974: 290; *Reninger v Fagossa* 1552). For example, in *Mouse's Case* (1608), a river barge was threatened by a storm. The passengers were in danger of drowning unless the barge's cargo was thrown overboard. A passenger started throwing cargo overboard. One of the items thrown was a box which belonged to a fellow passenger called Mouse. The box contained £113, and Mouse sued the passenger for the loss. However, the court found that the passenger was not liable for the damages, because he had acted out of necessity: whilst it would ordinarily be a criminal offence to throw Mouse's box overboard, the passenger was forced to choose between the competing interests of saving human lives or saving Mouse's box; and faced with this decision, the act of throwing Mouse's box overboard could not be regarded as unlawful.

I shall make necessity more precise by contrasting it with a related defence called duress of circumstances. The kind of choice required for a necessity defence – one between competing values or interests – often arises in concert with extreme circumstantial pressure. But circumstantial pressures are not relevant to the doctrine of necessity. This is because the necessity defence absolves liability by providing a *legal justification* as opposed to a *legal excuse*. Duress of circumstances holds that in some situations, circumstantial pressures are sufficiently extreme such that (i) the law considers the defendant's actions legally impermissible, but (ii) does not hold the defendant legally responsible for her actions. Hence, duress of circumstances provides a legal excuse. In contrast, necessity holds that sometimes a defendant will (i) perform an ordinarily criminal action, but (ii) their action is not criminal because they faced a decision which made it necessary to break the ordinary wording of the law. As Arnolds and Garland put it,

> [T]he courts limit the defence of duress to fear of serious bodily injury or death and make the defence personal to the person threatened. It makes no sense, however, to put those restrictions on the defence of necessity, since necessity is a justification and not an excuse (1974: 290).

Importantly, it is not required that the defendant faces a choice where *whatever* action she pursues violates the criminal law. Necessity can provide a defence in cases where the defendant breaks the law to prevent some greater evil occurring, even though the defendant would not have broken the law in letting the greater evil occur (*Re A* 2001). On a first pass, we can think of necessity as a *lesser of two evils* defence: it holds that a defendant is justified in breaking the law to prevent a greater evil occurring. Consider again *Mouse's case*. Whilst the disposal of Mouse's property is in general wrong because it violates Mouse's property rights, disposing of Mouse's property in the circumstances of the sinking barge is not wrong as Mouse's property rights could be justifiably infringed in order to save lives. As we shall see, Santoni de Sio formulates his reading of necessity in opposition to a rather extreme form of this lesser evil view, which he calls the *simple utilitarian reading* of the doctrine of necessity. On this view, legal necessity permits individuals to act so as to minimise harm in difficult situations. This view admits far too much. But it will serve as a good starting point for our discussion, as Santoni de Sio's analysis is in effect a sustained argument against this utilitarian reading of necessity.

3.3.2. Santoni de Sio's Analysis

Before I outline the central themes of Santoni de Sio's (2017) discussion of the legal doctrine of necessity, I shall make two clarifications.

First, Santoni de Sio's discussion is extensive. At times, it goes beyond the scope of the moral design problem. I shall not discuss the parts of his argument which pertain to property damage, as the discussion here is limited to interpersonal moral dilemmas involving harm to persons. Neither shall I discuss the issues arising out of (i) AVs being used as public service vehicles; (ii) AV manufacturers' duty of care to passengers; (iii) motor vehicle users' duty of care to road-users; and (iv) broader questions about whether AV collision algorithms should be regulated by a public authority. These considerations are important in determining the scope of the doctrine of necessity as a legal justification for harm and property damage caused in collisions. But the aim of Santoni de Sio's legal-philosophical approach is to answer the moral design problem, which is not concerned with these legal matters.

Second, Santoni de Sio's discussion of necessity is structured as follows. First, he discusses two landmark cases: *R v Dudley and Stephens* (1884) and *Re A* (2001). He argues that both cases place restrictions on the doctrine of necessity which render it incompatible with the simple utilitarian reading of the doctrine. This move is confusing because the doctrine in its present form was defined in these cases.[21] It was not defined in opposition to an earlier utilitarian doctrine of necessity. Next Santoni de Sio invokes some considerations to argue against the utilitarian reading of the doctrine. I shall focus on the most important: incommensurability and the right to life. Last, Santoni de Sio provides some practical recommendations based on his reading of the doctrine of necessity in answer to the moral design problem.

I start with the landmark cases. In *Dudley and Stephens*, four sailors were cast adrift off the coast of Africa. After many days without food, the cabin boy fell sick, and two of the sailors – Dudley and Stephens – decided to kill the cabin boy and eat him. The sailors were then rescued. When they returned to England, Dudley and Stephens were put on trial for murder. They attempted a necessity defence: if they had not killed the cabin boy, they would have starved to death. Lord Coleridge agreed that, had they not killed the cabin boy, they would have died. But he rejected this fact as sufficient grounds for a necessity defence. Santoni de Sio claims that Lord Coleridge's judgement was based on the principle that 'no innocent life should be taken under any circumstances' (2017: 415). This is almost certainly not true. Lord Coleridge's argument is more accurately construed as saying that no legal justification or excuse for killing the cabin boy obtained given the facts of the case.

Santoni de Sio then argues that 'things have changed' since *Dudley and Stephens*. He cites Brooke LJ's judgement in *Re A* as evidence for this. In *Re A*, the judges had to decide whether to permit a team of doctors to perform an operation that would separate conjoined twins. If the operation was not performed, both twins were guaranteed to die within a short time period. If the operation was performed, the weaker twin was guaranteed to die in the process. Brooke LJ permitted necessity as a legal justification for killing the weaker twin. He then set out three conditions for a successful necessity defence to killing in similar circumstances:

[21] To give a speculative diagnosis, I think that Santoni de Sio's treatment of necessity is confusing because he does not appreciate the distinction between deontological lesser-evil justifications and utilitarian justifications. The utilitarian thinks that we should save the greater number because what matters morally is aggregate welfare, and saving the greater number maximises aggregate welfare. In contrast, deontological lesser-evil justifications take as the object of aggregation what Seth Lazar (2012) calls *morally weighted harm*. The morally weighted harm inflicted on a person is the amount of harm inflicted minus the harm that they are liable to suffer given facts about their actions, their responsibility, and so on. I discuss this distinction in more detail in Chapter 5.

a) the person killed was here the one (involuntarily) impeding the survival of the other; b) the person killed would have certainly died anyways in a short time; c) the killing has been committed with the official permission of a public authority (Santoni de Sio 2017: 416).

Santoni de Sio's claim that 'things have changed' should be taken with a pinch of salt. This is because Brooke LJ's judgement in *Re A* is an extended argument in favour of the view that performing the operation is consistent with Lord Coleridge's judgement in *Dudley and Stephens*. But this has little bearing on Santoni de Sio's account. Santoni de Sio's next move is to argue against the simple utilitarian view of the doctrine of necessity by appeal to the judgements in these two cases. First, drawing on *Dudley and Stephens*, Santoni de Sio argues that the doctrine of necessity originally held that no innocent life should be taken in any circumstances. Second, Santoni de Sio claims that *Re A* relaxes this restriction, in that it allows killings that are consistent with (a)-(c). Because both requirements conflict with the simple utilitarian reading of necessity, Santoni de Sio concludes that the simple utilitarian reading is false. Santoni de Sio then tries to make precise exactly what it is that distinguishes the legal doctrine of necessity from the simple utilitarian view.

First, Santoni de Sio invokes what he calls the problem of incommensurability. He distinguishes three interpretations of this problem. On a *conceptual* reading, it is impossible to compare the value of different lives because there exists no objective metric for measuring the value of different lives. On an *epistemic* reading, it might be true that there exists an objective metric for measuring the comparative values of different lives, but agents forced to make the kinds of decision that the moral design problem is concerned with are unlikely to have the resources to make a sound evaluation. On a *normative* reading, it does not matter whether the comparative value of lives can be measured: each individual has a right to life, and this must be respected irrespective of how valuable that life is. So, what Santoni de Sio claims is incommensurable, conceptually, epistemically, or normatively, is the value of one person's right to life and the value of another person's right to life. Second, Santoni de Sio (2017: 419) argues for the primacy of the right to life over the right to be saved. Following Christie (1999), he argues that an innocent bystander's *right to life* has stronger force than an individual's *request* to be rescued.

Santoni de Sio (2017: 426) then concludes with the beginnings of an answer to the moral design problem:[22]

[22] I include three of Santoni de Sio's (2017: 428) nine points. The remaining conclusions reference the legal issues excluded from the discussion at the start of this section.

1. [there] might in principle be circumstances where a vehicle may be programmed to kill (*Re A*).
2. Given the strong restrictions to the intentional killing of innocents outside self-defence (*Dudley and Stephens*), the problem of the incommensurability of values and the right to life of persons, a program that allows for a vehicle to systematically hit persons who wouldn't be involved in the accident but for the vehicle decision seems unacceptable.
3. Based on the current legal constraints on killing under necessity (*Re A*), the intentional programming of an AV to target another AV might in principle be permitted under very specific and complex circumstances which seem unlikely to be realised in the near future.

3.3.3. The Dilemma

I now raise a dilemma against Santoni de Sio's answer to the moral design problem. In §3.2, I argued that the legal-philosophical approach admits two readings. I now argue that, on either reading, Santoni de Sio's answer to the moral design problem falls short of the aim of the legal-philosophical approach.

I start with the wide reading. The aim of the legal-philosophical approach, on this reading, is to use the law to resolve our disagreements about which moral principles provide the correct account of justified harm. But what Santoni de Sio has done is provide an alternative account of justified harm which is based on the doctrine of necessity. His account stands in competition with existing accounts, such as utilitarianism and contractualism. He has not explained *why* advocates of competing views will accept his alternative. Indeed, there is good reason to think that advocates of at least one view will *not* accept his necessity-based account. Consider utilitarianism. Santoni de Sio's necessity-based account of justified harm is unlikely to convince utilitarians for two reasons. First, in some cases, it delivers the wrong verdict: Santoni de Sio takes some harms as impermissible which utilitarianism deems permissible (e.g. the killing of an innocent person who is *not* involuntarily impeding the survival of another, in order to save a greater number of lives).[23] Second, the approach is justified by appeal to considerations which the utilitarian does not take as morally salient: the problem of incommensurability and the right to life. So, without an explanation of why the utilitarian should accept his

[23] Not all forms of utilitarianism permit such killings. But my argument succeeds provided there exists at least one plausible account of utilitarianism which would permit this.

competing account of justified harm, it is unclear how Santoni de Sio's necessity-based account can resolve our moral disagreements. In turn, it is unclear how Santoni de Sio brings us to a consensus on the moral design problem.

I now address the narrow reading. The aim of the legal-philosophical approach, on this reading, is to bring us to a consensus on the moral design problem *despite* our moral disagreements. The first problem is the same as above. Why should utilitarians accept Santoni de Sio's necessity-based answer to the moral design problem given that it appeals to principles which they do not recognise? I suspect Santoni de Sio would argue as follows: the law reaches its conclusions through a combination of moral and practical reasoning, which allows us to reach definite normative conclusions in the absence of any agreement about which moral principles give the correct account of justified harm (Santoni de Sio 2017: 413).

But the utilitarian can respond: the moral design problem is a *moral* problem. Answers to the moral design problem must be justified with moral reasons. Perhaps the legal-philosophical approach can bring us to a consensus on an *all things considered* design problem, which factors in both moral and practical reasons. But the aim of the legal-philosophical approach is to bring us to a consensus on the *moral* design problem despite our disagreements about which moral principles provide the correct account of justified harm. Santoni de Sio needs to provide a *moral* argument for his necessity-based solution to the moral design problem. Furthermore, advocates of competing accounts of justified harm must have reason to accept this argument despite their disagreements about which moral principles are correct. Without this argument, it is unclear how Santoni de Sio's necessity-based solution can bring us to a consensus on the moral design problem which advocates of different moral principles can accept *despite* their disagreements.

In §3.4 and §3.5, I attempt to salvage the narrow reading from this objection. The objection arises because Santoni de Sio provides an analysis of necessity based on the moral principles which underpin it. This is problematic because disagreement about moral principles gives rise to the problem which the legal-philosophical approach aims to overcome. I shall provide an alternative account of necessity, based on the *conditions under which* it provides a justification for inflicting harm, as opposed to the principles on which the justification is based. I argue that the conditions under which the doctrine applies are captured by a restricted Pareto principle. I use this principle to formulate a partial answer to the moral design problem. The answer is *partial* insofar as it provides an account of justified harm in some, but not all, of the collisions with which we are concerned. I argue that this

principle is agreeable from the perspectives of three different moral theories: utilitarianism, contractualism and deontology. If I am correct, then the legal-philosophical approach can bring us much closer to a consensus on a decision-rule to use as a partial answer to the moral design problem in the absence of any agreement about which moral principles are correct. So, whilst the law might not be able to bring us to a *complete* consensus on the moral design problem despite our moral disagreements, I believe it can bring us to a partial consensus.

3.4. The Restricted Pareto Principle

I now present an alternative reading of the doctrine of necessity. I then use this to formulate a *partial* solution to the moral design problem. I start with a statement of the principle which, I argue, captures the conditions under which necessity provides a justification for inflicting harm. According to

> *The Restricted Pareto Principle (RPP):* In situations where (i) harm to at least one person is unavoidable, and (ii) a choice about how to allocate harm between different persons is required, then *if* there exists a unique Pareto efficient allocation of harm, *then* bringing about the Pareto efficient allocation of harm is justified.

Three clarifications: First, an allocation of harm is Pareto efficient if, and only if, there exists no alternative allocation of harm in which all affected parties are at least as well-off, and some affected party is strictly better off. Second, there is a *unique* Pareto efficient allocation of harm when there is exactly one Pareto efficient allocation of harm among the alternatives. Third, RPP is a *sufficient* condition on justified harm in the cases where it applies. It does not follow from RPP that it is unjustifiable to bring about an outcome which is *not* uniquely Pareto efficient in scenarios where (i) and (ii) obtain.

I have explained what RPP amounts to. I now defend RPP as a plausible reading of the doctrine of necessity. My argument draws on the distinction that Brooke LJ drew between *Re A* (2001) and *Dudley and Stephens* (1884) in *Re A*.

Consider again the facts of *Re A*. Gracie and Rosie were conjoined twins. The doctors faced two options: (1) perform the operation and separate the twins; or (2) do not perform the operation. The medical evidence suggested that, on (1), the probability of Gracie surviving was 90% and the probability of Rosie surviving was 0%. On (2), both were almost certain to die. Brooke LJ permitted the defence of

necessity as a legal justification for the doctors' performing the operation and killing Rosie in the process. As Lord Coleridge denied necessity as a defence to killing the cabin boy in Dudley and Stephens, Brooke LJ needed to show that the facts of *Re A* were sufficiently different, such that the precedent in *Dudley and Stephens* did not preclude necessity as a justification for killing Rosie in *Re A*.

To understand Brooke LJ's distinction between the two cases, let us examine *why* Lord Coleridge disallowed necessity as a justification for killing in *Dudley and Stephens*. Consider,

> By what measure is the comparative value of life to be measured? Is it to be strength, or intellect, or what? It is plain that the principle leaves him who is to be profited by it to determine the necessity which will justify him in deliberately taking another's life to save his own. In this case the weakest, the youngest, the most unresisting, was chosen. Was it more necessary to kill him than one of the grown men? The answer must be "No" (1884: 287-8).

Lord Coleridge makes two points here. First, it is unclear how, if at all, we can measure the relative value of the lives of different persons. Second, that there was no good reason to kill the cabin boy instead of one of the other sailors.[24]

Brooke LJ distinguished *Re A* from *Dudley and Stephens* by considering these remarks. First, if Lord Coleridge had permitted necessity as a defence for killing the cabin boy, then the law would need to have weighed-up the comparative value of the sailors' lives against that of the cabin boy. However, in *Re A*, Rosie was guaranteed to die irrespective of whether the operation was performed. So, permitting necessity as a defence to killing in *Re A* would not require the law to weigh-up the comparative value of Gracie and Rosie's lives. Second, in *Re A*, the decision to kill Rosie instead of Gracie was not arbitrary: there was a good reason to kill Rosie, because she was impeding Gracie's survival, and Gracie was not impeding her survival. But in *Dudley and Stephens*, all the sailors were impeding each other's survival in the relevant sense: any one could have been killed to save the other two. Brooke LJ clarified his distinction by providing an analogy. Consider,

> [T]he same considerations would apply if a pilotless aircraft, out of control and running out of fuel, was heading for a densely populated town. Those in the aircraft are in any event "destined to die". The would be no question of human

[24] It should be noted that there was a reasonable expectation that the cabin boy would die sooner than the others because he had fallen sick. But it is clear from Lord Coleridge's judgement that this was insufficient reason to choose to kill the cabin boy over one of the other sailors.

choice in selecting the candidates for death [...] if their inevitable deaths were accelerated by the plane being brought down on waste ground (*Re A* 2001: 85).

There is a game-theoretic similarity between the facts of *Re A* and the aeroplane example, and this similarity is not shared by *Dudley and Stephens*.[25] Consider the situation of *Re A*, depicted in a payoff matrix:

		Rosie	
		Operate	Not Operate
Gracie	Operate	Rosie Dies Gracie Lives	Rosie Dies Gracie Dies
	Not Operate	Rosie Dies Gracie Dies	Rosie Dies Gracie Dies

The outcome if Gracie and Rosie both choose Operate is Pareto efficient, because there exists no other outcome which makes it the case that both twins are at least as well-off and at least one twin is strictly better off. All the remaining outcomes are Pareto inefficient. So, *Re A* had a *unique* Pareto efficient outcome, where both Gracie and Rosie 'agree' to perform the operation. In *Dudley and Stephens*, granting that at least one person had to be killed if any of the parties were to survive,[26] there is no unique Pareto efficient outcome. If Dudley and Stephens kill the cabin boy, this outcome is Pareto efficient, because there is no other outcome such that all three are at least as well off and at least one is strictly better off. The same holds if Dudley and the cabin boy kill Stephens, or if Stephens and the cabin boy kill Dudley.

Thus, the salient feature of Brooke LJ's distinction between *Re A* and *Dudley and Stephens*, is that *Re A* had a unique Pareto efficient outcome and *Dudley and Stephens* did not. I therefore take RPP as a plausible account of the conditions under which

[25] Note: I have ignored the probabilities in *Re A* here. Because the outcomes of performing or not performing the surgery were almost certain, I am treating them as certainties.
[26] The jury found that "there was no appreciable chance of saving life except by killing someone for the others to eat" (*Dudley and Stephens* 1884: 275).

necessity provides a justification for harm.[27] We can use this to formulate a partial answer to the moral design problem. Consider,

> *The Restricted Pareto Principle** (RPP*)*:* In collisions where (i) harm to at least one person is unavoidable, and (ii) a choice about how to allocate harm between different persons is required, then *if* there exists a unique Pareto efficient allocation of harm across different persons, *then* programming an AV to bring about the Pareto efficient allocation of harm is justified.

I shall now argue that RPP* is a decision-rule which at least three of our best moral theories can agree on as a partial answer to our question.

3.5. Overcoming Disagreement

I now argue that, in the scenarios where it applies, RPP* is a principle which utilitarians, contractualists and deontologists have reason to accept. Their reasons for accepting RPP* are grounded in the moral principles which they endorse.

I start with utilitarianism. I take it that most utilitarians believe that, in collisions where harm is unavoidable, AVs should be programmed to bring about the outcome which maximises utility (Bonnefon et al. 2016). In the collisions where it applies, RPP* justifies programming AVs to bring about an outcome only if that outcome is utility-maximising. This is because if one outcome Pareto-dominates another, then the dominating outcome has strictly greater utility than the dominated outcome (Coleman 1980: 515). As a unique Pareto efficient allocation of harm Pareto-dominates all other outcomes, it follows that the unique Pareto efficient allocation of harm is utility-maximising. So, the utilitarian has reason to accept RPP* because, on their view, AVs ought to be programmed to maximise utility in collisions, and RPP* does this in all the collisions where it applies.[28]

---

[27] It might be objected that the Restricted Pareto Principle (RPP) does not provide a complete account of the conditions under which the doctrine of necessity provides a justification for harm. This might be true, at least insofar as RPP does not consider the normative positions of the relevant parties such as who is responsible for the collision. This prima facie concern is addressed when I discuss deontology in the next section.

[28] Note that the sort of utilitarianism I consider here is maximising act utilitarianism. It might be objected that maximising act utilitarians would also consider the long-term consequences of accepting a rule like RPP*. Perhaps, if the public knew that under some conditions, AVs are programmed to kill their passengers, they would be reluctant to purchase AVs. As AVs are much safer than non-AVs, perhaps adopting a rule like RPP* would, therefore, cause more deaths in the long run. I am unconvinced. RPP* justifies programming the car to kill its passenger only if the passenger was going to die anyway. I doubt people would be reluctant to purchase AVs in light of this.

I now consider contractualism. I discuss T.M. Scanlon's (1998) contractualism, as this is plausibly the most sophisticated version of the view. Scanlon argues that the rightness of moral principles is determined by the justifiability of those principles to individuals affected by their prescriptions. An act is permissible, Scanlon argues, if it is permitted by any principle for the general regulation of behaviour that no one could reasonably reject. Roughly, an individual has grounds to reasonably reject a principle when it gives insufficient weight to her moral claims. These claims include considerations about her utility, her being treated unfairly, her not being accorded appropriate respect, and so on. What matters is that the considerations are *personal* to the individual in question, in the sense that the considerations have to be about how a proposed principle would treat her or would imply about her.

The contractualist has at least three reasons to accept RPP* as a partial answer to the moral design problem. First, an important feature of Scanlon's view is the *individualist restriction*, which holds that we cannot aggregate the moral claims of separate persons. The claims of each person must be assessed in isolation. RPP* determines that an outcome in a collision is justified by considering the utility of each affected person in isolation. For each person, a unique Pareto efficient outcome is such that either (i) the individual would rationally prefer that outcome to any other, as it causes the least harm to *them*; or (ii) the individual would not strictly prefer any alternative outcome, as no outcome will bring about strictly less harm to *them*. The aggregation of moral claims is not relevant to determining the unique Pareto efficient outcome, and in turn, it is not relevant to RPP*. Hence, RPP* is consistent with the individualist restriction.

Second, RPP* can be understood as a welfarist principle, insofar as it considers only the *utility* of affected persons conditional on each outcome. Scanlon admits considerations about utility as relevant to whether a principle can reasonably be rejected. But utility is not the only consideration. It might therefore be objected that RPP* is non-contractualist because it fails to account for non-welfarist considerations like fairness and responsibility. I find this objection unconvincing. In explaining why, I hope to show that RPP* captures another feature of Scanlon's view. Scanlon argues that:

> In many cases, gains and losses in well-being (relief from suffering, for example) are clearly the most relevant factors determining whether a principle could or could not be reasonably rejected (1998: 215).

I imagine that AV collisions where harm is unavoidable and there exists a unique Pareto efficient allocation of harm are cases where the welfare of each person is what matters most to the justifiability of the moral principles in question. At the very least, it strikes me that fairness and responsibility are not obviously relevant to whether RPP* can reasonably be rejected.

Scanlon (1998: 212-213) maintains that a person can reasonably reject a principle on the grounds of fairness only if the principle arbitrarily favours one person over another. RPP* justifies saving one person and killing another only if the individual killed would die irrespective of the AV's actions. In this respect, the decision to kill one person to save another is non-arbitrary, and RPP* cannot reasonably be rejected on the grounds of fairness. As for responsibility, I take it that *even if* a person is responsible for causing a collision, they can reasonably reject any principle which mandates that the AV should bring about her death, when this could be prevented at no cost (in welfare terms) to any other affected party.[29] So, it strikes me that a contractualist response to the moral design problem would take welfare as the principal moral consideration; and to this extent, RPP* captures another feature of the contractualist position.

Third, Scanlon endorses a principle which is stronger than RPP*. According to

> *The Rescue Principle:* [...] if you are presented with a situation in which you can prevent something very bad from happening, or alleviate someone's dire plight, by making a slight (or even moderate) sacrifice, then it would be wrong not to do so (Scanlon 1998: 224).

If welfare is the only relevant consideration in AV collisions where harm is unavoidable, then RPP* never requires an affected party in a collision to make a sacrifice. All parties would be at least as badly off on any other alternative. So, if the Rescue Principle is true, then programming an AV to bring about the same degree of harm to someone in a different way in order to save another affected party from death or serious harm is not something that can reasonably be rejected.

---

[29] It might be objected that RPP* can reasonably be rejected on *broader* grounds of responsibility. I have not considered the *incentives* that RPP* would create. Perhaps if RPP* were implemented in AVs, then reckless drivers in manual vehicles would be assured that AVs would act to save them from death, if this could be done without causing additional harm to other parties in the collision. So, plausibly AV passengers could reasonably reject RPP* on the grounds that it incentivises irresponsible or reckless driving. I cannot dispel this objection in its entirety. However, there is a slim chance of a *particular* reckless driver causing a collision with an AV, where this collision has a unique Pareto efficient allocation of harm *in the driver's favour*. So, we can at least say that reckless drivers do not have *good reason* to believe that RPP* would significantly improve their survival prospects should (i) RPP* be implemented into AVs and (ii) they continue to drive recklessly.

That concludes the contractualist argument. I now turn to deontology. Because deontology reflects a family of views, I cannot discuss all the considerations that deontologists would take as salient to the moral design problem. What I shall do instead is argue that certain features of RPP* are attractive to deontologists given the moral commitments that deontologists can in general be assumed to hold.

First, RPP* is a non-aggregative moral principle. The kind of aggregation that is important here is interpersonal aggregation. We can suppose that a principle is interpersonally aggregative just in case it combines morally relevant properties that pertain to separate persons into a single value (Hirose 2015: 24). RPP* is straightforwardly non-aggregative. This is clear from the fact that it satisfies Scanlon's individualist restriction. But at the risk of labouring the point, what is required for RPP* to provide a justification for bringing about an outcome is that *nobody* strictly prefers another outcome, and *somebody* strictly prefers this outcome. Hence insofar as deontologists are reluctant to accept aggregative principles, the fact that RPP* is explicable without reference to value impartially considered is a point in its favour for the deontologist.

Second, because RPP* is a welfarist principle, it might be argued that RPP* is insensitive to certain deontic considerations. First, consider *liability*. Roughly, a person is liable to be killed in self- or other-defence if, in virtue of posing an unjust threat to the life of another for which she is morally responsible, she forfeits her right not to be killed (McMahan 2009: 175-6, 2005; Firth and Quong 2012: 677). In the simplest case, Evil Murderer decides to kill Victim for no good reason, and Evil Murderer will kill Victim unless Victim kills Evil Murderer first. In this case, Victim is morally permitted to kill Evil Murderer in self-defence. Deontologists will standardly explain this permission by appeal to liability. Victim is permitted to kill Evil Murderer *because* Evil Murderer has forfeited his right not to be killed in virtue of posing an unjust threat to Victim's life for which he is morally responsible.

RPP* makes no reference to liability. This might be seen as problematic because deontological explanations of moral permissions in driving typically appeal to the concept of liability. Consider the following case from Jeff McMahan:

> *The Conscientious Driver:* A person who always keeps her car well maintained and always drives carefully and alertly decides to drive to the cinema. On the way, a freak event that she could not have anticipated occurs that causes her to veer out of control in the direction of a pedestrian (McMahan 2009: 165).

McMahan argues that the conscientious driver is liable to defensive force. If the pedestrian had a ray-gun, they would be permitted to vaporise the conscientious

driver to save their own life. McMahan's argument is, briefly, that the driver 'knows that driving is an activity that has a very tiny risk of causing great harm – so tiny that the activity, considered as a *type* of activity, is entirely permissible. But she has bad luck. [...] She has no positive reason to engage in the activity that she knows has a tiny risk of unintentionally killing an innocent bystander' (*Ibid.*).

Though RPP* makes no reference to liability, I shall argue that the concept of liability cannot be invoked to argue against RPP*. Consider,

> *Deliberate Head-On Collision:* Evil driver is driving on the wrong side of the road around a blind bend at exceptional speed. Evil Driver's intent is to kill other road-users even if she will die in the process. An AV is driving around the bend in the other direction. It contains one person, who we can call Passenger. The AV detects the other vehicle. It has two options. It can brake, in which case it will collide with Evil Driver's car, and both Passenger and Evil Driver will die; or it can swerve off the road, in which case Evil Driver lives and Passenger dies.

This collision has a unique Pareto efficient outcome. It is the outcome on which Passenger dies and Evil Driver lives. Hence RPP* implies that the AV is justified in killing Passenger and saving Evil Driver. The deontologist can reject RPP* on grounds of liability only if, in cases like *this*, considerations about liability imply that the AV is morally required to bring about the Pareto-dominated outcome. On two plausible views about liability, the AV lacks this moral requirement.

According to *internalism about liability*, 'If harming a person is unnecessary for the achievement of a relevant type of goal, that person cannot be liable to be harmed' (McMahan 2009: 5). This is McMahan's view. More precisely, McMahan claims that two conditions are required for liability to lethal defensive force. One is that the person is agentially responsible for some threat to the life of another.[30] The other is that killing the person is necessary to avert the threat. Because the threat to Passenger cannot be averted by killing Evil Driver, the internalist is forced to accept the conclusion that Evil Driver is not liable to defensive force (c.f. Unaicke 2014: 66; Firth and Quong 2012: 689-90). Hence the internalist cannot reject RPP* on liability grounds because Evil Driver is not liable to be killed on this view.

According to *externalism about liability*, what it means for Evil Driver to be liable is that a presumptive reason not to harm Evil Driver no longer applies in virtue of the fact that Evil Driver poses a threat to the life of Passenger. Hence on this view

[30] McMahan does not require the person to be *morally* responsible, only that the threat posed traces back to their agency. See Chapter 5 for an extended discussion of McMahan's view.

there is no positive moral reason to harm Evil Driver (Frowe 2014: 91). On the most plausible form of this view, 'The attacker retains [...] a humanitarian right to reasonable aid and protection, and this explains why the attacker can be both liable to defensive harm, and yet also be wronged by defensive harm that is unnecessary' (Firth and Quong 2012: 693). Here the externalist view implies that the AV is morally required to bring about the unique Pareto efficient outcome. This is because Evil Driver's humanitarian right to reasonable aid and protection provides a positive moral reason for Passenger or a third party to save Evil Driver's life if doing so imposes little or no cost on Passenger. As saving Evil Driver imposes no additional cost on Passenger, the externalist view agrees with RPP* that the AV is justified in bringing about the unique Pareto efficient outcome.

Because *Deliberate Head-On Collision* represents the strongest challenge that considerations about liability could press on RPP*, we should accept that liability poses no threat to RPP*. Hence the deontologist should accept RPP*, or at least not reject this principle on grounds of liability. A final point I should make is that the deontologist could in principle reject RPP* on grounds of desert. Whilst it may be true that considerations of liability provide no positive moral reason to bring about the Pareto-dominated outcome in *Deliberate Head-On Collision*, deontologists are free to claim that Evil Passenger *deserves* to die in virtue of posing the threat, and that considerations about desert justify the AV killing Evil Driver. In response, I am hesitant to accept that considerations about desert *ever* justify killing people. But even if they did, it seems morally inappropriate for AVs to take desert into account when making decisions. If it is true that Evil Driver deserves to die, then Evil Driver can be killed later after the facts have been settled in a fair hearing.

I argued that the deontologist cannot reject RPP* on the grounds that it is an aggregative principle or that it conflicts with considerations about liability. I also tried to dampen the criticism that RPP* conflicts with considerations about desert. Hence I hope to have shown that RPP* is a principle that is at least consistent with three of our best moral theories: utilitarianism, contractualism, and deontology. The arguments here are not conclusive. But I hope, at least, to have illustrated that the following claim is plausible: RPP* is a partial answer to the moral design problem which many of us can accept *despite* our disagreements about which moral principles offer the correct account of justified harm. This, I hope, constitutes a partial vindication of the narrow reading of the legal-philosophical approach.

3.6. Rejecting the Restricted Pareto Principle

I have tried to develop the most plausible account of what the legal-philosophical approach can offer. I am unconvinced that the law contains the seeds for a new moral theory that each of us has reason to accept. But I do believe that the law can reveal a point at which our three best moral theories converge. I have argued that RPP* is a condition under which these theories agree on what morality requires. Whilst our best moral theories can agree on RPP* as a partial solution to the moral design problem, I do not think that the legal-philosophical approach merits further consideration. RPP* lacks some important abductive qualities; and because RPP* is the most plausible consequence of the legal-philosophical approach, we should reject the legal-philosophical approach in favour of some other methodology.

The main problem is that RPP* applies in a narrow class of collisions. That is, it applies only in collisions where there exists a unique Pareto efficient outcome. Whilst there are some plausible collision scenarios where a unique Pareto efficient outcome exists, e.g. the Head-On Collision case, most collision scenarios do not have a unique Pareto efficient outcome. This suggests that RPP* falls far below our expectations for an answer to the moral design problem, as presumably it is at least a minimal requirement that the answer delivers a verdict in all or most collisions.

To compound the problem, there is no straightforward way to relax RPP* that retains its plausibility from the points of view of the different moral theories that I considered in the previous section. On the one hand, we can drop the requirement that the Pareto efficient outcome be *uniquely* Pareto efficient. On this view, the AV is justified in bringing about an outcome if that outcome is Pareto efficient. The convergence argument does not work for this view because some Pareto efficient outcomes are not optimific. Hence the support from utilitarianism is lost. It is only unique Pareto efficient outcomes that are guaranteed to maximise utility impartially considered.[31] On the other hand, the requirement for Pareto efficiency could be weakened to a requirement for Kaldor-Hicks efficiency. Here a Kaldor-Hicks improvement is one on which the individuals who gain could in principle hypothetically compensate those that are made worse-off, leading to a Pareto improvement (Kaldor 1939; Hicks 1939). An outcome is Kaldor-Hicks efficient if there is no Kaldor-Hicks improvement on that outcome. Though Scanlon's rescue

[31] It is true that given affected parties, $i = 1,2,3, \ldots, n$, each with a utility function u_i, then then for any Pareto efficient outcome, $x \in X$, there exists a vector of weights $\vec{a} = (a_1, \ldots, a_n)$, such that $a_i > 0$ for all i, such that x maximises $\Sigma a_i u_i(x_j)$. But there is no guarantee that each $a_i = 1/n$, which is what is required for utilitarianism (Negishi 1960).

principle implies that it is permissible to bring about *some* Kaldor-Hicks efficient outcomes, i.e. those where one person is made marginally worse-off to render another person significantly better-off, the Scanlonian contractualist would not endorse the view that it is *always* justifiable for the AV to act so as to bring about a unique Kaldor-Hicks efficient outcome (Scanlon 1998: 224). Hence RPP* cannot obviously be made more applicable without losing its argumentative support.

3.7. Conclusion

In this chapter, I developed Santoni de Sio's (2017) legal-philosophical approach to the moral design problem, and then rejected this approach. The aim of the legal-philosophical approach was to use the criminal law to formulate an account of what is morally required in AV collisions that overcomes our moral disagreements. Santoni de Sio's account of moral AV behaviour based on the legal doctrine of necessity did little in the way of resolving our moral disagreements. The reason for this is that Santoni de Sio's account was in effect a form of rights-based deontology that at least utilitarians have good reason to reject. I argued that a more plausible approach is to provide an account of the conditions under which the AV is justified in killing as opposed to that in virtue of which the AV is justified in killing. I then proposed the Restricted Pareto Principle as a condition on justified harm which utilitarians, contractualists and deontologists could plausibly accept. However, the problem is that the Restricted Pareto Principle applies in a small class of cases. There is also no obvious means to relax the principle whilst remaining agreeable from the point of view of the three moral theories that I considered. Hence we should reject this principle and in turn reject the legal-philosophical approach.

4. Rawls, Maximin and Leximin

4.1. Introduction

Suppose that an autonomous vehicle (AV) encounters a situation on the road where (i) imposing a risk of harm on at least one person is unavoidable; and (ii) a choice about how to allocate risks of harms between different persons is required (Lin 2016; Goodall 2014). What does morality require in these cases? How, morally, should AVs be programmed to allocate harm or risks of harm between the different parties? I call this the *moral design problem* (Keeling 2017, 2018a, 2020).

Many people endorse a utilitarian answer to the moral design problem (Bonnefon et al. 2016). According to this approach, AVs should be programmed to minimise expected harm or loss-of-life in collisions. Derek Leben (2017) recently proposed a contractualist alternative to the utilitarian approach. His answer is based on John Rawls' (1971) theory of justice. Whilst utilitarians are concerned with maximising some conception of *the good*, such as pleasure or wellbeing, contractualists are concerned with the justifiability of moral principles to those affected by their prescriptions (Scanlon 1998).

In this chapter, I argue that we should reject Leben's contractualist answer to the moral design problem. In §4.2, I explain the main ideas from Rawls' theory of justice which feature in Leben's answer to the moral design problem. In §4.3, I explain Leben's answer to the problem. In §4.4, I argue that Rawls offers less support for Leben's algorithm than we might initially expect. In doing so, I aim to show that Leben owes an independent argument for his answer to the moral design problem. In §4.5, I raise three objections to Leben's answer, all of which must be overcome if he is to provide a plausible defence of his view. In §4.6, I block a possible escape route for Leben. In §4.7, I conclude.

4.2. Rawls on Justice

In this section, I explain the central ideas of Rawls' (1971) theory of justice which feature in Leben's (2017) answer to the moral design problem. This account of Rawls is incomplete in many respects. But I hope it will provide sufficient grounding for the discussion.

I start with two general remarks. (1) Amartya Sen (2009: 5-7) distinguishes two methodological approaches to questions about justice. On the one hand, some philosophers have tried to pinpoint what is necessary and sufficient for a just society. On the other hand, some philosophers have developed comparative conceptions of justice; the aim being to stipulate a criterion by which social arrangements can be evaluated as *more just* or *less just* relative to one another. Rawls' theory of justice is an example of the first approach. His theory tells us what is required for a society to be just. (2) Rawls is part of the *social contract* tradition. The social contract theorists aim to ground the state's authority to restrict the freedoms of citizens in the actual, or hypothetical, consent of those citizens. In short, the state is justified in restricting the freedoms of citizens only if the citizens could at least hypothetically consent to those restrictions because it is in their interests to do so.

I now describe three features of Rawls' theory. First, Rawls' argues that justice is fairness. In Rawls' view, political society is a system of cooperation between individuals. The society is just when the terms of cooperation are fair to all those involved. Second, Rawls' main concern is with the principles of justice which regulate the *basic structure* of society. That is, the political and social institutions which 'assign basic rights and duties, and regulate the division of advantages that arise from social cooperation over time' (Rawls 2001: 10). Third, Rawls aims to provide both a method to determine the fair terms of cooperation and a statement of those terms. The method is a thought experiment called the *original position*, and the terms are Rawls' *two principles of justice*. I describe each of these in turn.

The original position is a hypothetical situation in which representative citizens decide on principles of justice to regulate the basic structure of society from a list of alternatives. These alternatives are taken to include things like utilitarianism, libertarianism, and the two principles of justice offered by Rawls (Rawls 1971: 122). Each party in the original position represents the interests of a sub-class of citizens; and all the citizens in society have a representative.

The parties in the original position must decide which principles of justice to adopt from behind a *veil of ignorance*, which Rawls describes as a situation where

'no one knows his place in society, his class, position or social status; nor does he know his fortune in the distribution of natural assets, his strength, intelligence, and the like' (*Ibid.*: 137). Rawls also states that the parties do not know whether the citizens they represent are in a minority or a majority. Neither do the parties know about the details of their life projects; their conception of the good; or their psychological dispositions such as risk-aversion, optimism or pessimism. Finally, the parties have no knowledge of the economic or political standing of their society, nor the level of civilisation or culture which the society can reasonably be expected to achieve (*Ibid.*: 137-142; Rawls 2001: 85-89).

The candidate principles of justice regulate the distribution of *primary goods* in the society, which are goods which all people have reason to want in order to facilitate their aims and ambitions. Examples include 'rights and liberties, powers and opportunities, income and wealth' (*Ibid.*: 62). Rawls assumes that all the parties in the original position are rational, and that each prefers more primary goods to less. So, whilst the representatives are unaware of their life aims and projects outside the original position, they have reason to select principles of justice which provide them with the freedoms and resources to pursue their life projects, no matter what these might be (*Ibid.*: 142-3). Rawls also stipulates that the parties in the original position are equal. In his words, 'the parties are equally represented [...] and the [principles of justice selected] are not influenced by arbitrary contingencies or the relative balance of social forces' (*Ibid.*: 120).

So, the original position is a bargaining situation which allows the parties to reach a *fair* agreement on principles of justice to regulate the basic structure of society. It does not presuppose an ethical or religious point of view in order to determine what is fair. Instead, it allows the parties to settle on mutually-advantageous principles of justice which benefit all citizens no matter what their life aims are or their conception of the good happens to be (Rawls 2001: 15). The agreement is fair for two main reasons. First, the veil of ignorance removes the unfair bargaining advantages which some individuals would ordinarily have over others in light of their social standing (*Ibid.*). Second, as the parties are ignorant of their position in society, they are unable to favour principles of justice because these principles confer benefits on *them* (*Ibid.*: 18; Harsanyi 1953: 434-5).

I shall presently turn to the principles of justice which, Rawls argues, the parties in the original position would agree upon. But first, I shall explain the decision procedure which Rawls believes the parties in the original position would use to discriminate between candidate principles of justice. According to Rawls, the

parties would appeal to the *maximin* decision procedure. This means that the parties would look at the primary goods available to the worst-off citizens under each of the principles of justice, and select principles which provide the greatest allocation of primary goods to the worst-off citizens (Rawls 1971: 150-161). The argument for maximin will be discussed in detail in §4. But the basic idea is that, as the parties in the original position do not know how the citizens they represent will fare under different principles of justice, they have good reason to favour principles which *guarantee* a minimal set of rights, liberties and opportunities for the worst-off citizens (Rawls 2001: 97-100).

I now turn to the principles of justice which, Rawls argues, free and equal citizens in the original position would agree upon. According to

> *The First Principle of Justice:* Each person has the same indefeasible claim to a fully adequate scheme of basic liberties, which scheme is compatible with the same scheme of liberties for all (Rawls 2001: 42).

According to

> *The Second Principle of Justice:* Social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least advantaged members of society (*Ibid.*)

Rawls argues that the first principle takes priority over the second. In short, the liberties of individual citizens cannot be compromised to bring about social or economic gains for society as a whole (Rawls 1971: 61). The first part of the second principle aims to ensure social mobility. The idea is that individuals are not constrained from improving their social standing in virtue of their initial social position. The second part is called the *difference principle*, and it holds that whilst the distribution of wealth and income need not be equal, inequalities in the distribution of wealth must be to the advantage of the worst-off (*Ibid.*).

This description of Rawls' theory does not come close to a comprehensive treatment of the ideas contained within Rawls' work. But I hope this description will provide sufficient grounding for the discussion in subsequent sections.

4.3. Leben's Answer

In this section, I explain Leben's answer to the moral design problem. Leben's answer uses two ideas from Rawls: the original position and the maximin rule.

First, Leben imagines that the affected parties in a particular AV collision enter a bargaining situation analogous to Rawls' original position.[32] The affected parties include *any* individual who could receive at least some harm conditional on at least one alternative available to the AV. The idea is that, in this hypothetical situation, the parties could reach a *fair* agreement on which alternative the AV should select. The parties are told the survival probabilities of each party conditional on the alternatives available to the AV; but they do not know which survival probabilities correspond to *them*. Furthermore, if multiple parties have the same survival probability on one of the alternatives, the parties are not made aware of this (which mirrors Rawls' decision to exclude information about the relative proportions of the population that each citizen in the original position represents).

Second, Leben contends that in these circumstances, the parties would choose between the AV's alternatives using an iterated form of the maximin rule, called *leximin*.[33] The leximin rule compares the survival probabilities of the worst-off person on each alternative (where the worst-off is understood as the individual with the lowest survival probability). It then selects the alternative which assigns the greatest survival probability to the worst-off person. In this respect, leximin is identical to maximin. However, the rules part company if the survival probabilities of the worst-off are equal on two alternatives. Maximin cannot discriminate between two such alternatives. In contrast, leximin compares the second-lowest survival probabilities on the remaining alternatives, and selects the alternative which gives the highest survival probability to the second worst-off person. If there is another tie, the third-lowest survival probabilities are compared, and so on. If two or more alternatives have identical profiles of survival probabilities, leximin randomises between them.

---

[32] Leben (2017) is, to my knowledge, the first person to apply the original position to the moral design problem. But it is worth noting that Thomas Shelling (2006) defends a similar position for determining whom to save in many-versus-one rescue cases.

[33] Leben does not use the term 'leximin'. He writes '[there] is one part of the Maximin procedure, that, to my knowledge, has not been worked out sufficiently well by Rawls or anybody else, and is perhaps the only original contribution that I have to make to the moral theory itself […] It seems clear that agents in the original position would also consider the *next-lowest* payoffs, since they have an equal chance of being the next player, and are interested in maximising her minimum as well' (2017: 110). The iterated form of maximin described by Leben is called leximin, and it has featured in moral philosophy (e.g. Otsuka 2006: 119-121; Hirose 2015: 29) and welfare economics (e.g. Sen 1976; Hammond 1975).

Based on this assumption about how the parties in his original position would decide between the alternatives in an AV collision, Leben develops a formal collision algorithm which uses the leximin rule. I shall illustrate Leben's algorithm with an example. Consider,

> *Example:* The AV can swerve left or swerve right. If the AV swerves left, Amy has a 60% chance of survival and Beth has a 30% chance of survival. If the AV swerves right, Amy has a 30% chance of survival and Beth has a 50% chance of survival.

Leben's algorithm first compares the survival probabilities of the worst-off parties conditional on each alternative. These are identical (30%), so the algorithm compares the survival probabilities of the second worst-off. In this case, 60% is greater than 50%, so the algorithm selects swerve left. If the survival probabilities for the second worst-off were identical, the algorithm would randomise between swerve left and swerve right.

This paper is critical of Leben's algorithm. I argue that we have reason to reject it as an answer to the moral design problem. But Leben and I agree about more than we disagree. I want to emphasise some of the excellent features of Leben's answer to the moral design problem before describing my criticisms.

First, Leben's answer is presented in decision-theoretic terms. It is not a set of moral principles, but an algorithm based on some principles. Leben therefore bridges the moral design problem with the related problem of how to programme moral principles into AVs. This is progress, as non-utilitarian moral principles are difficult to formalise. Furthermore, Leben's algorithm reaches a verdict in all the collisions with which the moral design problem is concerned. This is an important step forward for non-utilitarian answers to the moral design problem, which are difficult to capture in algorithms which cover all collisions which might arise.

Second, Leben's algorithm is based on contractualist principles. It therefore resists some of the objections which can be raised against utilitarian collision algorithms. It requires no interpersonal utility comparisons; and it does not demand that AV manufacturers sacrifice their passengers to save the greater number in collisions. At the heart of Leben's algorithm is the idea that AVs ought to be programmed to allocate harm in accordance with principles which are justifiable to the recipients of harm. It is impermissible, on Leben's view, to impose burdens on some to confer benefits on others. The maximin principle is designed to bring about Pareto efficient allocations of harm. (An allocation of harm is Pareto efficient if, and only if, there exists no alternative on which everyone is at least as well-off, and

someone is strictly better-off.) So, whilst the moral motivation for Leben's project might at first be unfamiliar, the motivations are nevertheless laudable. To this end, my criticisms of Leben are not intended as an indirect defence of a utilitarian answer to the moral design problem. Instead, I hope these criticisms can be used to develop an even stronger non-consequentialist answer to the moral design problem.

4.4. A Rawlsian Algorithm?

Leben (2017: 108) views his collision algorithm as a straightforward application of Rawlsian principles to the moral design problem. In this section, I challenge Leben's assumption that his collision algorithm is defensible on Rawlsian grounds.

There is an important disanalogy between Rawls' original position and Leben's original position for AV collisions. In Rawls' case, the parties are deciding between alternative principles of justice which regulate the distribution of primary goods in society. Rawls writes:

> I shall simply take as given a short list of traditional conceptions of justice [...] I then assume that the parties are presented with this list and required to agree unanimously that one conception is best among those enumerated (Rawls 1971: 122).

According to Rawls, the parties would appeal to the maximin principle when deciding between the principles of justice, and in doing so favour his two principles over utilitarianism (*Ibid.*: 175-183). Importantly, the parties in Rawls' original position are *not* choosing between alternative distributions of primary goods using the maximin principle. I believe that Leben has misread Rawls here. He writes:

> [...] the original position is a method limited to determining the distribution of what Rawls calls 'primary goods' (Leben 2017: 109).

In Leben's original position, the parties must decide between alternatives in a particular AV collision using information about the distribution of survival probabilities associated with each alternative. The alternatives, in effect, represent possible distributions of a single primary good, namely the survival probabilities of each affected party (*Ibid.*). Leben argues that the parties in his original position would use the maximin principle to decide between the alternative distributions of this primary good. Leben's original position is therefore analogous to a variation

on Rawls' original position in which the affected parties decide between competing distributions of primary goods using the maximin principle.[34]

This disanalogy between Rawls' and Leben's accounts of the original position makes Leben's answer to the moral design problem hard to defend on Rawlsian grounds. Rawls and Leben are defending different claims. Rawls says that the parties deciding between principles of justice would, in the original position, decide between those principles using the maximin principle (Rawls 2001: 97). Leben says that parties deciding between distributions of survival probabilities would, in the original position, use maximin to decide on one such distribution. I shall develop this point into two criticisms of Leben's answer.

The first criticism concerns the veil of ignorance. Leben restricts the information available to the parties in his original position. I do not think these restrictions are defensible on Rawlsian grounds.

Rawls developed the original position to provide hypothetical circumstances under which free and equal citizens can reach a fair agreement on principles of justice to regulate the basic structure of society. The veil of ignorance is imposed for two reasons: it removes unfair bargaining advantages and it restricts the kind of arguments that might be provided in favour of some principles over others (*Ibid.*: 81-8). In short, Rawls does not want parties in the original position to select principles of justice *because* those principles favour individuals in *their* social circumstances (*Ibid.*; see also Harsanyi 1953). However, one of the most important features of the original position is that the parties have sufficient knowledge and understanding to evaluate the alternative principles of justice on offer (including, at least, utilitarianism and the two principles offered by Rawls). No information is excluded about what these principles entail.

The same cannot be said about Leben's veil of ignorance. Some information is excluded about the alternatives on offer. The parties are given information about the distributions of survival probabilities on each alternative. But if the same survival probability applies to two or more persons, the parties are not made aware of this (Leben 2017: 112). So, Leben's veil of ignorance does more than remove unfair bargaining advantages and prevent arguments of the form 'we should choose alternative x because it is in *my* best interests'. It also provides an incomplete

[34] A further disanalogy is that whilst *survival* is a primary good, it is not obvious that the probability of survival is a primary good. So, the parties in Leben's original position are choosing between alternative gambles concerning a primary good.

description of the alternatives available to the parties. In some cases, this is problematic. Consider,

> *Problem Case:* The AV can swerve or continue its current path. If the AV continues, there is a 100% chance that its passenger and four pedestrians will sustain a fatal injury. If the AV swerves, there is a 100% chance that its passenger will sustain a fatal injury.

The parties in Leben's original position would know only that at least one person is guaranteed to die on either alternative. Leben argues that, with this limited information, the parties would agree to use his leximin algorithm and randomise between the alternatives. The parties do not have sufficient information to perform a utilitarian calculation, as they do not know which alternative maximises total or average survival probabilities over all affected parties. They also lack the requisite information to argue in favour of swerve on the grounds that swerve Pareto-dominates continue. (On swerve, at least one party is better-off, and no party is worse-off.) Because the parties have such limited information, I agree with Leben that leximin is the most sensible rule to use to discriminate between the alternatives.[35] But I agree with Leben only because the parties have insufficient information to use *any other* plausible algorithm for discriminating between the alternatives. It is unclear *why* the parties should be unaware of the fact that five people stand to die on one alternative, and just one of these same people on the other. Adding this information would not provide unfair bargaining advantages to the parties in Leben's original position. Neither would it allow them to favour one alternative over the other because *they* stand to be better-off on that alternative.

There are two conclusions to draw here. First, Rawls' argument for imposing the veil of ignorance in the original position does not support the restrictions on information which Leben employs in his analogue of the original position for AV collisions. In Rawls' original position, the parties have sufficient information to evaluate the alternatives on offer. This is not so in Leben's case. Second, Leben in effect gerrymanders his veil of ignorance to ensure that the affected parties are forced to decide between the alternatives using his leximin algorithm. There exists no other rational decision procedure which could be used. This strikes me as a misunderstanding of Rawls' motivation for using the veil of ignorance. Rawls did not want to *force* the parties to accept his two principles of justice by restricting the

[35] Note that, with complete information, leximin mandates saving the greater number in many-versus-one cases (Hirose 2015: 164-5). So, Leben advocates using leximin *given the information available*, but leximin would not mandate randomising if complete information about the survival probabilities were given.

information available to them. He wanted to show that, under fair bargaining conditions, it is in the parties' rational self-interest to accept his principles of justice.

I now turn to the second criticism, which concerns the maximin principle. It is worth noting from the outset that Rawls did not defend maximin as a universally applicable decision-rule. He writes:

> [...] the maximin rule was never proposed as a general principle of rational decision in cases of risk and uncertainty, as some seem to have thought [...] Such a proposal would be simply irrational [...] The only question is whether, given the highly special and unique circumstances of the original position, the maximin rule is a useful heuristic rule of thumb for the parties to organise their deliberations (Rawls 2001: 97).

Rawls (2001: 97-8) argues that maximin is a plausible decision rule to use in the original position only if certain conditions are met. (1) The parties in the original position must have no knowledge of the probability of an arbitrary citizen being represented by each party in the original position. In other words, the parties in the original position do not know whether the citizens they represent are in a minority or a majority. (2) Because maximin concerns the worst-off citizens under different principles of justice, it must be necessary for the parties to be significantly more interested in the primary goods which can be guaranteed, as opposed to those which can be gained. Rawls argues that this condition is satisfied to some degree when the guaranteeable level for each citizen is 'quite satisfactory'; and that it is fully satisfied only if the guaranteeable level is 'completely satisfactory'. (3) As maximin examines only the worst-case scenario on each alternative, it must be the case that the worst outcomes on all other alternatives fall substantially below the guaranteeable level.

The first condition presents a trade-off for Leben. On the one hand, Leben can restrict the information available to the parties in his original position, such that the parties have no knowledge of the number of individuals who correspond to each survival probability. In this case, the first condition is satisfied. But the parties are unable to adequately evaluate the alternatives available to them in the collision, which I have argued is an undue restriction on the information available to them. On the other hand, Leben can relax the restriction on information, and allow the parties knowledge of *how many* people correspond to each survival probability. But then Rawls' first condition for maximin is not satisfied. So, Leben satisfies the first condition for maximin only if he imposes an undue restriction on the information available to the parties in his original position.

The second condition is fully satisfied in Rawls' original position when the guaranteeable level for each citizen is 'completely satisfactory' (*Ibid.*). If this condition obtains, Rawls argues that citizens will be more interested in securing basic rights and opportunities, rather than focusing on the additional rights and opportunities which could be gained. In the circumstances of an AV collision, the guaranteeable level is unsatisfactory. These collisions are such that death or serious harm is very likely for at least one affected party, and a choice about how to allocate harms across different parties is required. I suspect the affected parties in such a collision would be more concerned with maximising their chances of survival, rather than guaranteeing a *completely satisfactory* survival probability. So, the second condition for maximin is not satisfied.

The third condition states that the worst outcomes on all other alternatives falls substantially below the guaranteeable level. Recall that for Rawls, the alternatives are principles of justice as opposed to distributions of primary goods. What Rawls has in mind here is that the worst-case scenario for a citizen (in terms of rights and opportunities) under a utilitarian principle of justice is substantially worse than the rights and opportunities which can be guaranteed under Rawls' principles of justice. We can imagine, for example, a situation where one class of citizens is enslaved for the benefit of another class of citizens. Rawls thinks that maximin is a justifiable decision rule to prevent this kind of scenario arising. In Leben's case, this third condition is not met. A utilitarian collision rule does not offer a *substantially worse* outcome for the worst-off affected parties than Leben's leximin algorithm. It is likely that, on either algorithm, the worst-off affected parties will be exposed to serious risks of death or harm.

So, Rawls provides three conditions which are jointly sufficient to warrant the use of maximin as a decision rule in the original position. None of these conditions are satisfied in the context of the moral design problem. It seems, therefore, that Rawls' argument for using maximin does not apply.

4.5. Objections to Leben's Algorithm

I have argued that Leben's (2017) answer to the moral design problem cannot be defended on Rawlsian grounds. This provides insufficient reason to reject Leben's answer. But it does motivate the need for Leben to provide an independent argument in favour of his answer. I now describe three challenges that must be overcome if Leben is to provide a defence of his view.

Challenge 1: In some collisions, Leben's answer mandates programming AVs to select alternatives which the affected parties could not rationally consent to, provided their preferences satisfy the von Neumann-Morgenstern (1953) axioms for rational preferences.[36]

The problem arises because Leben evaluates the alternatives in collisions using survival probabilities. He assumes that 'injuries like broken ribs, whip-lash, etc., can be represented as points along the dimension of likelihood of survival' (Leben 2017: 111, 113). Survival probabilities are therefore intended as a proxy for physical harm. Leben acknowledges that a one-dimensional scale of this kind is not entirely plausible, as some non-fatal injuries might be considered equivalent to or worse than fatal injuries. But irrespective of what kind of injury is placed at the bad end of the scale, Leben's decision to evaluate alternatives using the probability of *worst-case scenario* injuries obtaining is what gives rise to our challenge. Consider,

> *Scenario 1:* The AV can swerve left or right. If the AV swerves left, there is a 0% chance that its passenger will sustain a fatal injury and a 100% chance that its passenger will sustain a lifelong debilitating injury. If the AV swerves right, there is a 1% chance that its passenger will sustain a fatal injury and a 99% chance that its passenger will remain unharmed.

Leben's algorithm selects swerve left because it gives the passenger the greatest chance of survival. I contend that there exists at least one scenario (equivalent to or analogous to *Scenario 1*) in which Leben's algorithm mandates programming the AV to select an alternative which the passenger could not rationally consent to. As mentioned, I shall assume that the passenger has rational preferences insofar as her preferences satisfy the von Neumann-Morgenstern (1953) axioms. I also assume that the passenger strictly prefers no injury to a lifelong debilitating injury; and that she strictly prefers a lifelong debilitating injury to a fatal injury. Note, however, that the passenger's preference ordering can be changed, and the same objection will arise.

Here is the problem: Leben's algorithm mandates swerve left no matter how low the probability of a fatal injury is on swerve right. It could be 1%, or 0.1% or 0.01%.

---

[36] The axioms: let $\prec$ denote strict preference, $\sim$ denote indifference and $\preccurlyeq$ denote weak preference. *Completeness* holds that for any two lotteries A, B, either $A \prec B$, $B \prec A$ or $A \sim B$. *Transitivity* holds that if $A \preccurlyeq B$ and $B \preccurlyeq C$ then $A \preccurlyeq C$. *Continuity* holds that, if $A \preccurlyeq B \preccurlyeq C$, then there exists a probability $p \in [0,1]$ such that $[pA + (1-p)C] \sim B$. *Independence* holds that if $A \prec B$, then for any C and $p \in [0,1]$, $[pA + (1-p)C] \prec [pB + (1-p)C]$. My argument makes use of the Archimedean Property, which is sometimes assumed instead of continuity. But if either continuity or the Archimedean Property is assumed, the other is entailed by the von Neumann-Morgenstern Expected Utility Theorem.

One requirement of von Neumann-Morgenstern rationality is that an agent's preference ordering is held fixed under sufficiently small deviations in probabilities. This is called the *Archimedean Property*. Formally, letting $\prec$ denote strict preference, the requirement is that for any lotteries A, B and C, if $A \prec B \prec C$, then there exists a small probability, ε, such that $[(1-\varepsilon)A + \varepsilon C] \prec B \prec [\varepsilon A + (1-\varepsilon)C]$.[37] As there is no ε, such that *if* the probability of a fatal injury on swerve right is equal to ε, *then* Leben's algorithm mandates swerve right, it follows that there exists at least one collision (equivalent to or analogous to *Scenario 1*), where Leben's algorithm mandates programming the AV to select an alternative which is not in the passenger's rational self-interest. Hence, in some collisions, Leben's algorithm mandates programming the AV to select alternatives to which the passenger could not rationally consent.

This is problematic for two reasons. First, many would argue that, in cases like these, it is morally permissible to take the gamble on behalf of the passenger and programme the AV to swerve right. If the gamble did not pay off, it would be a reasonable moral justification to highlight that we programmed the AV to select the option which we rationally expected to bring about the best outcome *for the passenger* (Otsuka 2012). Second, I take it that Leben's algorithm is intended to be an algorithm to which affected parties in collisions could at least hypothetically consent. Unless Leben assumes that affected parties in collisions have *irrational* preferences, at least by von Neumann and Morgenstern's (1953) standards, it seems that in at least some collisions, the affected parties could not rationally consent to Leben's algorithm.

This challenge arises because Leben, in effect, uses the maximin principle twice in his algorithm. First, the algorithm evaluates each alternative based on the probability of a *worst-case scenario* obtaining for each affected party. Second, the algorithm selects the alternative which provides the best deal for the worst-off party. The objection can therefore be avoided by removing the first use of maximin. For example, we might instead calculate the expected utility for each affected party conditional on the alternatives; and then select the alternative which gives the greatest expected utility to the worst-off party. This preserves the primary usage of maximin, whilst making it at least somewhat plausible that the affected parties in a collision could rationally consent to Leben's algorithm.

[37] The lotteries in square brackets should be read, e.g. 'A with a probability $1-\varepsilon$ and C with a probability ε'.

Challenge 2: The maximin rule gives undue weight to the moral claims of the worst-off. Consider,

> *Scenario 2:* The AV can swerve left or right. If the AV swerves left, there is a 100% chance that its passenger will die, and twenty nearby pedestrians will be unharmed. If the AV swerves right, there is a 99% chance that its passenger will die, and a 100% chance that twenty nearby pedestrians will receive lifelong debilitating injuries.

Leben's algorithm selects swerve right. Indeed, Leben's algorithm selects swerve right no matter how many pedestrians stand to receive lifelong debilitating injuries. Leben (2017: 144) acknowledges this counterintuitive feature of his algorithm and offers two points in response. First, he argues that scenarios of this kind are unlikely to arise. But the fact that these scenarios are *unlikely* does nothing to address the moral complaints of the pedestrians when such scenarios do arise. Leben's second response is that '[he] would always prefer to be one of the injured pedestrians (and [he] would thus prefer the action which produces the minimum outcome)' (*Ibid.*). So, whilst many pedestrians might receive lifelong debilitating injuries, no one *individual* receives an injury worse than the fatal injury which the passenger would sustain on swerve left. I am unconvinced. First, Leben's algorithm is meant to be contractualist: it aims to programme the AV to allocate harm in a way that is justifiable to each affected party. If we added a twenty-first pedestrian to *Scenario 2*, this would make no difference to the calculation performed. *Prima facie*, the moral claims of the twenty-first pedestrian are not given due consideration, because for Leben's algorithm, it is irrelevant whether or not she is present in the collision (Hirose 2015: 74). Second, I would prefer to lose a limb rather than die. It does not follow that, if forced to choose between killing one person and removing a limb from every human on the planet, I have stronger moral reasons to choose the latter option. The fact that I would *prefer* to lose a limb rather than die is not a good moral reason to inflict a very large number of serious injuries to prevent a single death (Norcross 1997).

Challenge 3: Suppose that Leben is correct about survival probabilities, and that he is also correct about maximin. In some collisions, there is another algorithm which assigns a higher survival probability to the worst-off than Leben's algorithm. Consider,

> *Scenario 3:* The AV can swerve left or swerve right. If the AV swerves left, there is a 0% chance that Anne will survive, and a 70% chance that Bob will survive.

If the AV swerves right, there is a 1% chance that Bob will survive, and a 60% chance that Anne will survive.

Leben's algorithm mandates programming the AV to swerve right. This is because the worst-off party on swerve right has a 1% chance of survival, and the worst-off party on swerve left has a 0% chance of survival. Leben's algorithm, then, assigns a survival probability of 1% to the worst-off party in *Scenario 3*.

I now introduce a rival algorithm, which we can call *greatest equal chances*. Leben contends that sometimes the AV ought to randomise between the two alternatives. So, there are at least three options in *Scenario 3*: swerve left, swerve right and construct a fair lottery between the alternatives. Plausibly, if the AV can construct a fair lottery, then it can also construct a weighted lottery. On the greatest equal chances algorithm, the AV is programmed to construct a weighted lottery between the alternatives, where the weightings are fixed to ensure that the affected parties receive the greatest equal survival probabilities. The process is akin to tossing a biased coin to decide whether to swerve left or right, where the degree to which the coin is biased ensures that Anne and Bob are given the greatest equal chances of survival. If $x \in [0,1]$ is the probability of swerve left in the weighted lottery, and $1 - x$ is the probability of swerve right, then the AV gives Anne and Bob an equal chance of survival provided $0.7x + 0.01(1 - x) = 0.6(1 - x)$. Solving for x, we see that if the AV assigns a probability of 0.456 to swerve left and $1 - 0.456$ to swerve right, then Anne and Bob have equal survival probabilities of 32.6%.

So, the greatest equal chances algorithm gives the worst-off a survival probability of 32.6%, which is greater than the 1% survival probability which Leben's algorithm assigns to the worst-off. If the alternatives in *Scenario 3* should be evaluated using survival probabilities; and if maximin is the rule that we have best reason to use, then we ought to adopt greatest equal chances in place of Leben's algorithm.[38] It follows that under Leben's two assumptions, there exist collisions in which another algorithm ought to be used to decide between the alternatives.

[38] The earliest statement of the relation between maximin and greatest equal chances is, to my knowledge, due to Derek Parfit (2003: 76-8). For discussions of lotteries like the one described see Rasmussen (2011), Rivera-López (2008) and Hirose (2015: 121-2).

4.5. Blocking Leben's Escape

One move that Leben could make in response to this third criticism is to adopt the greatest equal chances algorithm. The last thing I shall do is block this move. The view that assigning the greatest equal chance of survival is a *fair* strategy for deciding who to save in interpersonal moral dilemmas is not new. Consider,

> *Lifeboat:* You are in a lifeboat in the midst of a storm. There are two rocks up ahead in the distance. On Rock A, there is one person who is stranded. On Rock B, there are five people who are stranded. For some reason, it is feasible to save only the one person on Rock A or the five people on Rock B. It is impossible to save everyone. There are no morally significant differences between the parties. Hence no person is responsible for the calamity or has a stronger moral claim to be saved, and there are no special ties between you and any of the parties.

What is the morally right way to decide who to save? Most people judge that the right thing to do is to save the greater number. But when we try to account for this judgement, we find ourselves appealing to claims about aggregate harm. We want to say that it is *worse* that five should die rather than one. John Taurek (1977) famously rejected this line of argument. According to Taurek, the right thing to do is to toss a coin to decide whether to save the one or save the five. Tarek's argument has three parts (Hirose 2015: 111-2). It is worth unpacking the argument in full.

First, Taurek argues that it is *permissible* to save the one rather than the five. To motivate this claim, Taurek (1977: 300-01) asks us to imagine that a person, David, owns a drug, and that David needs this drug to survive. However, there are five strangers that David could save by giving them the drug instead of keeping it for himself. Taurek contends, I think rightly, that David is permitted to keep the drug for himself. The next move that Taurek makes is suspect. He claims that *because* David is permitted to keep the drug for himself, that a morally motivated stranger forced to choose between giving the drug to David or to the five others, would be morally permitted to give the drug to David rather than to the five strangers. This move is suspect because, as Iwao Hirose (2015: 114) remarks, 'the permissibility of an act is not transferable from one person to another.' Presumably, David is morally permitted to keep the drug for himself because morality is not so demanding that it requires us to sacrifice *ourselves* to save *others* if we are not morally responsible for the situation. This reason does not transfer to the morally motivated stranger.

Second, Taurek argues that there is no impersonal sense in which it is worse that five should die rather than one. My death is bad *for me*, but not bad *tout court*.

This is a strange and deeply confused claim. First, if Taurek's point is taken at face value, then it would be unproblematic to save nobody at all. This is because at least one person is certain to die irrespective of whether we save the person on Rock A or the people on Rock B. If we save nobody, then six people will die. But we do not have the resources to say that six deaths is worse than one death, because each person's death is bad *for them*, but not bad *simpliciter* (c.f. Hirose 2015: 116).

Second, even if we deny that there exists a coherent impersonal perspective, this does not preclude our taking an impartial standpoint. Presumably, what Taurek takes issue with is the moral relevance of the mysterious perspective that Henry Sidgwick (1907: 420) called the *point of view of the universe* (c.f. Nagel 1986). What is strange about Sidgwick's point of view is that it takes the distribution as the basic unit of axiological significance, rather than the individual (c.f. Bader ms). This does not sit well with the intuition that, e.g. being happy, is good *for people* and not *good simpliciter*. However, we might argue that a perspective on which the right thing to do is invariant under permutations in the identities of the parties is the appropriate moral standpoint. Here we are not invoking any impersonal value over and above what is prudentially good for each person. Instead, the suggestion is that morality requires us to broaden the scope of our concern from *our* prudential value to that of *all* the affected parties. Accordingly, we could then appeal to an anonymous Pareto principle to argue that saving the five is better than saving the one. Consider,

(Person 1 Dies, Person 2 Lives, Person 3 Lives, Person 4 Lives, Person 5 Lives)

(Person 1 Lives, Person 2 Dies, Person 3 Dies, Person 4 Dies, Person 5 Dies)

If the parties did not know *who they were*, then presumably they would agree that the upper distribution is better than the lower distribution (Otsuka 2006: 121-23). Thus the incoherence of the impersonal point of view does not on its own rule out the adoption of an impartial standpoint which justifies saving the greater number.

Third, Taurek's final move is to argue that, because our substantive moral aim is to show equal respect to each of the affected parties, we ought to toss a coin to decide which rock to travel to. He writes: 'Each person's potential loss has the same significance to me, only as a loss to that person alone. Because, by hypothesis, I have an equal concern for each person involved, I am moved to give each of them an equal chance to be spared from his loss' (Taurek 1977: 307). Tossing a coin gives each person the greatest equal chance of survival – a 50% chance – and so, Taurek concludes that the morally right thing to do is toss a coin to decide who to save.

In response to Taurek's third point, Frances Kamm (1993: 101, 114-19) and Thomas Scanlon (1998: 232-33) have argued that greatest equal chances fails to show equal respect to each party. This is because, if we toss a coin, then each of the five has a moral complaint: their presence made no difference to our decision. Had there been four, three, or two, people on Rock B, our decision would have been the same. The coin would be tossed irrespective of how many people were on Rock B, provided there is at least one person on Rock B. This argument is not watertight. As Mike Otsuka (2006: 114) points out, if the presence of the additional people made *no difference*, then if the coin lands in favour of Rock B, we would go to Rock B, rescue one person, and leave the others behind. Otsuka's criticism is, I think, a little unfair to Kamm and Scanlon. For Kamm and Scanlon, the object of *fairness* is the decision-method. Given that the natural foil to greatest equal chances is a biased coin-toss that allows the people on Rock B to pool their chances together into a 5/6 chance of survival, it remains true that the unbiased coin-toss makes no difference to the decision-method (Kamm 1993: 148-9; see also Broome 1998: 960). It seems to me that fairness is more plausibly interpreted as demanding a 5/6 chance for the people on Rock B, rather than the 1/2 chance that Taurek proposes.

What does all this tell us? First, the argument for greatest equal chances is not great. Hence there is a strong presumption against using this algorithm to decide the allocation of risks of harm in AV collisions. But it gets worse. The argument for greatest equal chances crucially depends on the stipulation that there are no special ties between the affected parties, and that each person has an equal moral claim to be saved. This condition is not satisfied in AV collisions. It may be true that one party is morally responsible for the collision; or that AV companies have special obligations to prioritise the safety of their passengers (or the safety of other road-users). It may well be morally significant whether the AV's passenger poses a lethal threat to another or is the subject of a lethal threat. Accordingly, *even if* the greatest equal chances algorithm is plausible in cases where there are no morally significant differences between the parties, this supposition is not met in the cases that we are concerned with here (c.f. Nyholm and Smids 2016: 1282-84). Hence I think that there are strong reasons for Leben not to endorse greatest equal chances.

4.6. Conclusion

In this chapter, I argued that we have reason to reject Leben's (2017) answer to the moral design problem. First, I argued that Rawls' (1971) arguments for the veil of

ignorance and the maximin principle do not support Leben's application of these tools to the moral design problem. In doing so, I established that Leben owes an independent argument in favour of his view. Second, I argued that Leben's algorithm is based on two problematic assumptions: (i) that we ought to evaluate the alternatives in AV collisions using survival probabilities; and (ii) that we ought to use the maximin principle to choose between the alternatives. I then argued that *even if* these assumptions are granted, there are some collisions in which a greatest equal chances algorithm is preferable to Leben's algorithm, because it provides a higher survival probability for the worst-off party. Third, I argued that we ought to reject the greatest equal chances algorithm because it is insensitive to morally salient considerations (such as who is morally responsible for the collision).

5. The Deontological View

5.1. Introduction

When are automated vehicles (AVs) morally permitted or morally required to kill or harm road-users? In what circumstances are AVs permitted to prioritise the lives of their passengers over the lives of pedestrians or other road-users? Which moral principles explain these constraints? My aim here is to answer these questions.

The view I defend is that the morality of AVs reduces to the moral principles that govern killing and harming in self-defence. The AV is permitted to kill or harm a road-user under the circumstances if, and only if, and because, its passenger is permitted to kill or harm that road-user in self-defence under the circumstances.

What does this entail? First, there is a strong presumption against killing and harming road-users. It is almost never permissible to kill or harm road-users. The exception cases involve road-users posing unjustified threats to the lives of the AV's passengers; or road-users who indirectly but culpably contribute to these threats. The one remaining exception is where the AV loses control due to a hardware or software malfunction, or bad road-conditions, and the AV must distribute risks of harm between its passengers and other road-users. In these cases, provided the passenger is not negligent or reckless in creating the circumstances that facilitate the AV's loss of control, the AV ought to minimise expected aggregate harm.

These claims will be clarified and qualified in due course. For now, I will make four points to set the scene.

The first point is that my argument is abductive. I aim to show that my view does a better job than its rivals at explaining our considered moral judgements in the cases with which we are concerned; and that it coheres with our best theories of permissible killing in other domains. I think that these two abductive virtues are the most important when deciding between rival theories of killing for AVs.

The second point is that I will provide a criterion of rightness and not a decision-procedure (c.f. Keeling 2020: 301-3). What this means is that I care about the question: 'What *makes* the AV's acts right?'. I am not asking: 'How should the AV deliberate in collisions?' (Brink 1986: 421; see also Bales 1971; Crisp 1992; Hooker 1990). The point of a criterion of rightness is to serve as an ethical standard to adjudicate between competing AV decision-making algorithms. As Seth Lazar (2018: 862) puts it, a criterion of rightness 'picks out the option that [the AV] should choose from any given option set.' It does not explain how the AV ought to identify the morally best option. However, even this may be too strong, as a given criterion of rightness may assign an ineliminable role to individual judgement in moral decision-making (c.f. Hooker 1996: 534-35; see also Kramer 2009; Nagel 1979; Shafer-Landau 1994; Ross 1930). The essential feature of a criterion of rightness is that it stipulates the right-making properties of the AV's acts and shows how these properties feature in explanations of what is right.

Third, I plan to argue from a deontological perspective.[39] I have no interest in formulating the morality of AVs as the implications of one fundamental moral principle. Like C. D. Broad, I think that all 'attempts to do this seem quite plainly to oversimplify the actual situation' (Broad 1930: 283-4; c.f. Ross 1930: 24; Nagel 1979: 131-2; Urmson 1975: 115). As far as I am concerned, attempts to reduce the requirements of morality to a single moral principle invariably involve overstating the explanatory force of that principle; or blindsiding a great deal of what matters morally (c.f. Ross 1930: 24; see also Kagan 1992: 226; Hooker 2000: 104–7). Instead, I adopt the standard approach in the ethics of killing, and try to build a systematic picture of the morality of AVs through a careful analysis of the moral positions of the affected parties in road-traffic scenarios that we can expect AVs to encounter (Frowe 2014: 2-6; Lazar 2009: 704-5; Tadros 2011: 4-8; Thomson 1990: 32-3).

Fourth, I shall assign significant weight to my intuitions about what is morally permitted in particular cases. I am not proposing to use Frances Kamm's (1993, 2007) method of trying to formulate a moral principle that explains my intuitive judgements across a broad set of imagined cases (c.f. Hurka 2016: 139). I agree with John Broome (1998: 958-9) that 'Kamm's arguments are powerless against people who do not share her intuitions, or are willing to give up their intuitions when they

[39] The flavour of deontology that I defend is loosely in the tradition of Immanuel Kant, in the minimal sense that, as Victor Tadros (2011: 2) puts it, it 'grounds constraints on moral action in respect for persons.' However, I think the origins of deontological theories in the ethics of killing are better located in the Roman Catholic ethical tradition. In particular, the doctrine of the double effect (c.f. Foot 1967/2002). Thus St Thomas Aquinas' discussion of defensive killing in *Summa Theologica* (II-II, Qu. 64, Art 7) is the correct point of origin.

have reason to do so' (c.f. Norcross 2008). Instead, I will use Judith Thomson's (1990: 32-3) method. What this means is that I will assign significant weight to intuitive moral judgements about which I am confident, and that I believe others will share. But all intuitive judgements are subject to revision if they do not square with the theory that on balance offers the best systematisation of our considered judgements about what morality requires in AV collisions (c.f. Rawls 1971: 48-51).

The stage is set. I proceed as follows. In §2, I introduce the deontological model that I plan to use. This is a rights-based model. Because a big part of my argument is that my own view coheres with our best theories of killing in other domains, I shall present the deontological view by appeal to cases from several different areas. Then I will focus in on AVs, and explain what the view looks like in this context.

In §3, §4, and §5, I examine some cases. I am aware that AVs will almost never encounter dilemmatic choices of the sort that philosophers indulge. In most cases, as Rebecca Davnall (2020) points out, the AV can just brake. The cases that I talk about avoid this solution. First, I discuss cases in which the AV's passenger faces a threat. It can avoid the threat only through evasive action. But there is a road-user in the way. Should the AV's passenger bear the threat? Or is the AV permitted to kill or harm the road-user and in doing so save its passenger? Second, I discuss cases involving reckless jaywalkers. When is the AV morally required to undertake dangerous evasive action to avoid reckless jaywalkers? When is it permitted to kill or harm the jaywalker? Finally, I discuss cases in which the AV loses control due to bad road-conditions or a hardware or software malfunction. How should the AV distribute the risks of harm between its passengers and other road-users? I know that these cases do not exhaust the space of dilemmatic collisions that AVs might find themselves in. But they are among the most difficult and most important cases.

In §6, I tie things together into a set of general principles, and argue that my theory fares better than its rivals in an abductive comparison.

5.2. Rights-Based Deontology

5.2.1. The Big Picture

Robert Nozick begins *Anarchy, State, and Utopia* with the claim that 'Individuals have rights, and there are things that no person or group may do to them (without

violating their rights)' (1974: xi). Like Nozick, I take rights as the central moral concept. Like Nozick, I do this because I think that the part of morality that I am concerned with is best modelled as including side-constraints (1974: 29-30f).[40]

The rights-model is particularly well-suited to the ethics of AVs for two reasons. First, teleological models, i.e. models which hold that an act is right if, and only if, it is optimal or maximal with respect to some moral utility function, take as the object of analysis that which the AV morally *ought to do*.[41] In contrast, rights-based models take as the object of analysis that which the AV morally *ought not do*. The starting point is that the AV is free to do what it likes, and then side-constraints are built in as restrictions on permissible conduct. Accordingly, the rights-based model can provide principles for the general regulation of AV behaviour without micromanaging the AV's acts across all situations that it might feasibly encounter. The rights-based model is therefore more economical. Second, AV decisions involve a trade-off between moral and prudential goals. The AV must balance its moral goal of road-user safety against its prudential goal of time-efficient driving. Though it is in principle possible to develop an all-things-considered utility function that adjudicates between the AV's moral and prudential goals, this is quite difficult (c.f. Keeling et al. 2019; and Chapter 6). On balance, it is far easier to explain the general moral constraints on AV behaviour, rather than what AVs ought to do in all cases.

I shall now make precise the concept of rights that I plan to work with. First, the legal theorist Wesley Newcomb Hohfeld (1919: 38) said that 'those who use the word and the conception "right" in the broadest possible way are accustomed to thinking of "duty" as the invariable correlative.' On Hohfeld's view, X has a right *against* Y that *p* if, and only if, Y has a duty to X that *p*. When rights are understood like this, they are called *claim rights* (Thomson 1990: 39-45; McMahan 2009: 62-3).

Three remarks: (1) Rights and duties describe moral constraints on our actions. I take rights as the explanatory concept. Duties obtain *in virtue of* rights. What I am assuming here is natural. If I tried to murder Tim, Tim might protest '*You ought*

[40] Nozick was open to morality being modelled in other ways. He says that: 'Unfortunately, too few models of the structure of moral views have been specified heretofore, though there are surely other interesting structures … An array of structures must be precisely formulated and investigated; perhaps some novel structure will then seem most appropriate' (1974: 29-30f; see also 45-6, esp. 46f). Note that others who favour a model-based picture have reached divergent conclusions about the best model, e.g. John Broome (1991: 19-20)

[41] See Broome (1991: 1-20; 2004b: 30-48) for discussions of teleological models. Whether a teleological model is optimising or maximising depends on the formal features of the moral betterness relation that the utility function represents. Optimal choice sets contain all the options that are at least as good as any other, and presuppose that the betterness ordering is reflexive, complete and acyclic (Sen 1971: 16). Maximising choice sets contain all the options that are not worse than any other, and accordingly do not require completeness.

not do that!'. If I ask *'Why?'*, Tim can respond *'Because I have a right not to be killed.'* Tim does not have a right not to be killed because I ought not kill him. (2) Y's duty that *p* is discharged if, and only if, *p* is true. There is no further requirement that Y makes it the case that *p* (Thomson 1990: 39f). (3) Because propositions stand in certain entailment relations to one another, so too do rights. Accordingly, if X has a right against Y that *p*, and *p* implies *q*, then X also has a right against Y that *q*.

What rights do we have? There is no need for a list. What matters here is that people have rights against all other people that those people do not kill or harm them. These rights ground a strong presumption against killing and harming. The strength of this presumption needs to be emphasised. As Lazar (2009: 701) tells us, 'The right to life is uniquely important and arguments for its diminishment or loss should meet a heavy burden of proof.' There are, I think, two justifications that can override the presumption against killing and harming others (Rodin 2011: 74-5). The first is what is called a *liability justification*. This is where a person has no moral complaint against being killed or harmed because of something they have done or failed to do (McMahan 2009: 8). The second is what is called a *lesser-evil justification*. This is where a person's rights against being killed or harmed can be justifiably infringed to prevent a substantially greater calamity occurring. What I shall do in the next two sections is explain how liability and lesser-evil justifications work.

5.2.2. Liability Justifications

In this section, I will explain the mechanics of liability justifications. First, I like to think about liability justifications as normative explanations. Consider,

The Structure of Liability Justifications:

P1 For all people *X* and *Y*, such that in general *X* has a right against *Y* that *p*, if *X* and *Y* satisfy conditions C, then *X*'s right against *Y* that *p* is suspended.

P2 X and Y satisfy conditions C.

C Therefore, X has no justified moral complaint against Y that *p*.

The *explanandum*, or the sentence describing the phenomenon to be explained, is a contingent normative fact about one agent lacking a justified moral complaint towards another in a particular set of circumstances (Hempel and Oppenheim 1948: 152). For example, 'Bob has no justified moral complaint against Alice that Alice flattens his legs with a steamroller.' The *explanans*, or the set of sentences intended

to account for the explanandum, contains: (i) a principle describing some conditions under which one person's right against another is suspended or absent; and (ii) a set of boundary conditions which relate that principle to the particulars of the case. The principle might be something like 'For all people, *X* and *Y*, *X's* right against *Y* that *Y* does not harm *X* is suspended if *Y* poses a lethal threat to innocent people.' The boundary conditions might take the form 'Bob poses a direct lethal threat to James and Sandra, and James and Sandra are innocent people.' Accordingly, liability justifications are explanations of why one person lacks a justified moral complaint against another by appeal to a general principle about the suspension of rights.[42]

The reason that I like to formulate liability justifications like this is that it allows us to systematise a bunch of applied ethical concepts into a single framework. What I shall do is present the three kinds of liability justifications: consent, forfeiture, and desert, as attempts to describe sets of conditions C to fit into the model I described.

The first kind of liability justification is where one person *waives* a right against another through *consent*. Roughly, P1 in consent justifications has the form: 'For all people *X* and *Y*, *X's* right against *Y* that *p* is suspended if *X* consents to *Y* that *p*' (Manson and O'Neill 2007: 72-77; Tadros 2016: 201-222; Thomson 1990: 349-73). To illustrate, Patient has a right against Nurse that Nurse does not stick a needle in Patient's arm. Thus Patient has a justified moral complaint against Nurse if Nurse sticks a needle in Patient's arm. But if Patient consents to Nurse sticking a needle in his arm, then Patient's right against Nurse is waived, and Patient has no justified moral complaint against Nurse if Nurse sticks a needle in his arm. Consent and its relation to liability are familiar from different departments of moral life. As George Fletcher (1996: 109) explains, 'When individuals consent to undergo medical operations, to engage in sexual intercourse, to open their homes to police searches, or to testify against themselves in court, they convert what otherwise would be an invasion of [...] their rights into a harmless or justified activity.'

Obviously, the devil is in the details. There are difficult questions of formulation that I have not addressed in this picture (Tadros 2016: Ch 10-14). For example, can I waive *any* right? It is one thing to consent to a nurse sticking a needle in my arm. It is quite another for me to consent to someone cannibalising me[43] or causing me

[42] The way that I have formulated liability justifications is based on Carl Hempel and Paul Oppenheim's (1948) deductive-nomological model of scientific explanations. On one hand, the explanation is deductive in that the explanans deductively entails the explanandum. On the other hand, the explanation is nomological in that it contains a non-redundant 'covering law.' In Hempel and Oppenheim's model, the covering law is a non-accidental empirical generalisation. In our case, it is a general moral principle (c.f. Salmon 1999: 14-17).

[43] See Friedman and Arold (2011) for a discussion of the German legal case that involved Armin Meiwes killing and cannibalising a consenting victim who he met on the internet.

severe bodily injuries.[44] Presumably, there are *some* restrictions on the rights that I can waive. Yet these restrictions cannot straightforwardly be understood in terms of a threshold, such that one can give consent only to harms of a particular severity. Presumably, in general one cannot consent to being killed. But, as Jeff McMahan (2002: 463) argues in the context of voluntary euthanasia, 'if it would be prudentially rational for a person to kill himself and his death would not be worse for others, it should be permissible for a third party to kill him at his request.' This seems right. But it suggests that the validity of consent to death or severe harms depends crucially on the presence of other factors. I shall return to this problem later on when I discuss the morality of AVs killing jaywalking pedestrians.

I now turn to the second kind of liability justification. This is when a person is liable to be killed or harmed because they have forfeited their rights not to be killed or harmed (Tadros 2011: 198-202; Thomson 1990: 361-73). Suppose that Medieval Enthusiast attacks Victim with a broadsword. Medieval Enthusiast intends to kill Victim. He will succeed unless Victim kills him first. Here Medieval Enthusiast poses an unjust threat to Victim. The threat is unjust because Victim's right not to be killed by Medieval Enthusiast will be violated if Medieval Enthusiast succeeds. Presumably, Victim is permitted to kill Medieval Enthusiast in self-defence. We can make sense of this moral permission by appeal to forfeiture. Because Medieval Enthusiast poses an unjust threat to Victim's life, he is liable to be killed by Victim.

There is a dispute about the mechanics of forfeiture. Three camps have emerged. According to the *Rights-Violation View*, Medieval Enthusiast forfeits his right not to be killed because, if Victim does not kill him, Medieval Enthusiast will violate one of Victim's sufficiently important rights (Thomson 1991). According to the *Moral Responsibility View*, Medieval Enthusiast forfeits his right not to be killed because he poses an unjust threat to Victim for which he is morally responsible (Otsuka 1994). According to the *Agential Responsibility View*, Medieval Enthusiast has forfeited his right not to be killed because he is agentially responsible for the threat to Victim's life (McMahan 2005, 2009; Frowe 2014). What it means for Medieval Enthusiast to be agentially responsible for the threat is roughly that the threat can be traced back to Medieval Enthusiast's actions, and those actions were

[44] The Law in England and Wales imposes limits on the degree of harm that one can consent to. For example, in *R v Brown* [1994] 1 AC 212, the House of Lords considered the question of whether consent to harms can provide a defence to actual or grievous bodily harm inflicted in the course of sadomasochistic sexual activities (c.f. s20 and s47 Offences Against the Person Act 1861). Lord Templeman ruled that 'Society is entitled and bound to protect itself against a cult of violence. Pleasure derived from the infliction of pain is an evil thing. Cruelty is uncivilised. I would answer the certified question in the negative.' For related cases see *R v Emmett* [1999] All ER (D) 641 (CA) and *R v Wilson* [1996] 3 WLR 125.

performed voluntarily, non-coercively, and under conditions of minimal rationality (Frowe 2014: 74–75; McMahan 2001: 401; see also Lazar 2009: 706).[45] Note that on the agential responsibility view facts about moral responsibility can play a role in determining the degree of harm that it is proportionate to inflict on the agent. But moral responsibility is not a necessary condition to be liable for defensive force.

The final ground for liability justifications is *desert*. We often take it for granted that certain people *deserve* to have harms inflicted upon them, or to have their property taken away from them, given acts that they have done or omitted to do. Indeed, as Victor Tadros (2011: 1) points out, 'not only do we typically think that it is acceptable to punish offenders, we think that states that fail to punish enough are unjust.' When a person deserves to be harmed, they are liable to be harmed. But desert-based justifications differ from forfeiture-based justifications. Roughly, forfeiture-based liability justifies harms that are *instrumental* to averting a threat. In contrast, desert-based liability justifies harms that are *non-instrumental* responses to acts that people have done or failed to do (McMahan 2009: 8, 2005: 1; Tadros 2016: 67–8; but see Firth and Quong 2012). I shall not have anything further to say about desert. But I think it is helpful to draw the contrast with forfeiture.

That concludes the introduction to liability justifications. Though people have rights not to be killed or harmed, in some cases, these rights are suspended due to consent, forfeiture, or desert. Both consent and forfeiture will feature heavily in the account of permissible killing for AVs that I shall later go on to provide.

5.2.3. Lesser Evil Justifications

I now turn to lesser-evil justifications. What is central to liability justifications is that a person's right not to be killed or harmed is suspended or absent. Lesser-evil justifications obtain in cases where a person retains their right against being killed or harmed. The idea is that in certain cases a person's right may be justifiably infringed to avoid a greater evil occurring (Rodin 2011: 74–5). Consider an example.

[45] Interestingly, in *Hill v Baxter* [1958] 1 QB 277, the defendant was charged with dangerous driving under s.11 of the Road Traffic Act 1930. The defendant had no memory of the episode, and submitted to the court that he was incapable of forming the intention to drive. Because under the 1930 Act, dangerous driving is a strict liability offence, the denial of intention was irrelevant (strict liability offences have no *mens rea* component). However, the defendant then attempted an automatism defence. He argued that the *actus reus* had not been performed because he failed to meet the standards of minimal agential responsibility. But unfortunately for the defendant, Lord Goddard ruled that he had simply fallen asleep.

Suppose that Red Shirt Killer intends to kill all and only those people dressed in red shirts. I am wearing a red shirt, and Red Shirt Killer is chasing me with an axe. In order to get to safety, I must cross a narrow bridge. You are on the bridge. You are safe because you are wearing a blue and green polka dot shirt. Because you are listening to music, you do not hear me cry, '*Get out of the way! I need to come through!*'. The bridge is not very high, and if I push you into the river below, you will at most sustain minor injuries. Suppose that I push you off the bridge in order to get past. Presumably, I did not act wrongly in doing so. However, you have not consented to the minor injuries, nor have you forfeited your right against being harmed, and nor do you deserve to be harmed. How then do we explain my moral permission? The standard answer is that your right not to be harmed was *justifiably infringed*, insofar as my pushing you off the bridge was the lesser of the two evils.

How do lesser-evil justifications differ from liability justifications? I will make four points. (1) Liability justifications are insensitive to aggregative considerations. As Helen Frowe (2014: 78) remarks, 'If a group of ten people set about kicking Victim to death, they are not each liable to one-tenth of lethal defensive force. Rather, each is liable to lethal defensive force.' In contrast, lesser-evil justifications are aggregative. Though it may be impermissible to kill one innocent person to save the lives of two others, presumably it is permissible to kill one innocent person to save the lives of two billion others. (2) The object of aggregation for lesser-evil justifications is not harm *simpliciter*, but instead what Lazar (2012) calls *morally weighted harm* (c.f. Frowe 2014: 67-70; Tadros 2011: 253). The morally weighted harm that is inflicted on a person is the total amount of harm inflicted minus the harm that the person is liable to suffer. To illustrate: If we believe that a person is liable only to defensive force that is minimally necessary to avert the threat which they pose, then any harm over and above that threshold is morally weighted harm. Hence whilst lesser-evil justifications can appeal to facts about aggregate harm, the aggregation is over morally weighted harm as opposed to harm *simpliciter*.

(3) When people suffer harms for which they are liable, they are not entitled to compensation. If Nick tries to kill me with a plant pot, and I break his arm in self-defence, I do not need to compensate Nick for the harm that I have caused him. The same is not true for lesser-evil justifications. If breaking Nick's arm is the only way to stop Plant Pot Killer from killing everyone in the garden centre, then if I break Nick's arm, I may owe him compensation or an apology for doing so. After all, Nick still retains his right against me not to be harmed; and I have infringed that right. (4) There is a close relationship between *proportionality* and lesser-evil justifications. McMahan (2011: 153-4) distinguishes between *narrow* and *wide* proportionality.

First, narrow proportionality is what is proportionate relative to the harms that a person is liable to suffer. If I inflict more harm on you than you are liable to suffer, then my act is narrowly disproportionate. Second, wide proportionality refers to what is proportionate in the context of a lesser-evil justification. For example, there may be some cases in which it is permissible to bomb innocent civilians in war, but the harm done to the civilians would be widely disproportionate if it were excessive.

5.2.4. Deontology for Automated Vehicles

I have explained the big picture. I shall now explain how the rights-based model can be translated into the context of AV decisions. Here is the principle that forms the backbone of my account of permissible killing for AVs. Consider,

> **The Self-Defence Principle:** The AV is morally permitted to kill or harm a road-user if, and only if, and because, its passenger is morally permitted to kill or inflict comparable harms on that road-user in self-defence.[46]

Four clarifications: (1) The Self-Defence Principle is a placeholder. I have not said what the self-defensive permissions of the AV's passenger are. So, the principle is silent on the content of the AV's moral permissions. (2) You can accept the Self-Defence Principle and deny the theory of self-defence that I provide. You can also accept the theory of self-defence that I provide and deny the Self-Defence Principle. (3) So, *my job* is to offer a plausible account of the AV passenger's self-defensive moral permissions; and to offer a principled reason to accept that the AV's moral permissions track these self-defensive permissions. (4) I shall assume for simplicity that the AV contains a single passenger. Because the moral permissions of different passengers inside the AV will be more or less the same, this idealisation is harmless.

The main argument for the Self-Defence Principle is abductive. I hope to show that this principle fares better than its rivals at explaining our considered moral judgements in the cases with which we are concerned. But I want to provide some initial motivation for the Self-Defence Principle before getting into the details.

First, the Self-Defence Principle is most plausible if we accept a deflationary picture of AVs and their moral status. I believe that the only moral agents in AV

[46] In a recent paper, Antti Kauppinen (forthcoming) argues that an AV is permitted to kill only if a human driver in the same circumstances is permitted to kill. This is roughly the same idea. But Kauppinen assumes this principle. He does not attempt to argue for it. Note that the Self-Defence Principle is equivalent to the principle that the AV's moral permissions are identical to those of a morally motivated stranger acting in other-defence.

collisions are the people involved in those collisions. AVs are just tools that their passengers use to achieve certain aims. Though AVs are *autonomous*, in the sense that these robots can set and pursue goals within certain well-defined parameters, the AV's passengers are ultimately the agents in control at all times on the road (c.f. Nyholm 2018c). Roughly, on this picture, AVs can be thought of as *extensions* of their passengers' agency as opposed to independent moral agents. Thus it makes sense to pin the AV's moral permissions to the moral permissions of its passenger.

One further commitment that renders the Self-Defence Principle plausible is the idea that AV collisions are morally unexceptional situations. What I mean by this is that there are no morally significant differences between AV collisions and other moral dilemmas that involve distributing risks of death or harm across multiple parties with conflicting interests. I discussed in Chapter 2 several arguments to the effect that AV collisions are in certain respects morally exceptional (Nyholm and Smids 2016; Himmelreich 2018; Goodall 2016). Whilst I think that there is much to be learned from these arguments, I think that AV collisions are exceptional only insofar as a great deal of work is required to discern the morally relevant facts. I am sceptical that the moral permissions of AVs are governed by moral principles that are in some sense fundamentally different to those that govern killing in other domains. There is, of course, room for reasonable disagreement on this point. But I hope to make clear that the ordinary moral principles that govern killing in self-defence are sufficient to capture everything that we need for the morality of AVs.

I will conclude this section by making precise two objections to the Self-Defence Principle that I cannot respond to in full. The first is that the Self-Defence Principle is *individualistic*, in the sense that it holds that the moral permissions of AVs can be settled by appeal only to facts about the individuals involved in collisions. I can see that this might give rise to the following concern: AV collisions are more complex than normal road-traffic collisions. This is because the AV's *manufacturer* or *designer* can in certain cases be morally responsible for death or harm caused by the AV. For example, if the AV kills a pedestrian because of a software error that ought to have been ironed out in the quality assurance process, then it seems that the Self-Defence Principle will give us an incomplete picture of the moral status of the AV's action. This is an important criticism. One point that I can make in response is that these issues are not unique to the context of AV collisions. It might be true that a person's car malfunctions and that this malfunction is a proximate cause of a collision. That these situations are possible is registered in existing accounts of self-defence that have touched on the moral permissions of drivers (McMahan 2009: 162-182; see also Frowe 2014: 80-85, 138-9; Lazar 2009: 711-727; Tadros 2011: 46-7, 181-86). I

shall do what I can to provide a more complete response to this objection later on in my analysis of cases in which AVs kill or harm people because of malfunctions.

The second objection is that a deontological criterion of rightness of the sort that I am proposing fails to take seriously the epistemic limitations of AVs. What is at issue here is that certain features of driving scenarios that I take to be morally relevant are unknowable given the AV's limited information about its environment (Nyholm and Smids 2016; Himmelreich 2018). This is again an important worry. I provide a more detailed treatment of AV decision-making under uncertainty when I discuss the risk-imposition problem in Chapter 6. But for the time being, I shall take for granted a subjectivist criterion of rightness, according to which the AV's moral permissions are indexed to its evidence about the morally relevant facts (c.f. Frowe 2010; Gibbard 1990; Kagan 1989; Scanlon 2008). This means that I shall judge the AV relative to its reasonable predictions about what is happening in its environment. Often, however, I shall talk in objective terms about the AV's moral obligations, and then refine these obligations given the AV's epistemic constraints.

What follows in the next few sections is a discussion of various scenarios in which it is unclear what the AV morally ought to do. I shall use the Self-Defence Principle and my views about the self-defensive permissions of the AV's passenger to work out what the AV morally ought to do. Before I get into that discussion, it is worth flagging that the Self-Defence Principle creates a strong presumption against the AV killing *anybody*. Accordingly, what morality requires in the vast majority of cases is that the AV drives with sufficient caution, such that it can come to a stop should something unexpected happen, e.g. a child running into the road, or a cyclist falling off their bike. I shall provide a treatment of the appropriate degree of caution for AVs to exercise in Chapter 6. But I hope to emphasise that the situations that I discuss here are quite exceptional. In general, AVs should behave so as to avoid killing or harming people unless extremely rare conditions obtain.

5.3. Obstructer Cases

Consider,

> *Massive Truck:* There is a massive truck headed towards the AV. Passenger will be killed in a collision with the truck. The only way to avoid the collision is for the AV to swerve into Motorcyclist, who would be killed by the impact.

Cases like this are called *Obstructer Cases*. What is at issue is whether the AV is morally permitted to kill Motorcyclist in order to avoid a lethal threat to Passenger given that Motorcyclist is not directly involved in the threat (Frowe 2014: 24-30; see also McMahan 2004; Otsuka 1994; Rodin 2002; Thomson 1991). Obviously, the particulars of the case are unimportant. The massive truck is a stand-in for any lethal threat that might befall Passenger, be it an out-of-control vehicle, or logs that have fallen off a truck, or a giant boulder that becomes loose in a landslide. The same is true for Motorcyclist. He is a stand-in for any road-user that is *in the way*, in the sense that the road-user blocks the AV's escape from the threat. The salient question is whether it can ever be true that the AV is permitted to kill a road-user that is in the way in order to avoid an independent lethal threat to its passenger.

These cases matter for two reasons. First, there are relatively few situations that require AVs to make life-or-death decisions. In most cases, the AV can brake to avoid a collision (c.f. Davnall 2020). What is distinctive about obstructer cases is that the AV cannot brake to avoid a collision, as this will result in Passenger's death. The choice is between allowing Passenger's death or killing someone else. I think it plausible that these cases will arise in practice. Often collisions can be avoided through evasive action. Often people get in the way. Second, I know that in practice AVs will not have full information about the collision. I like to think of obstructer cases as limiting cases of more complex situations in which the AV must either allow its passenger to receive a serious risk of death or harm; or impose a serious risk of death or harm on another road-user. I hope that these limiting cases are instructive about more complex cases involving the distribution of risks of harm.

Obstructer cases have already been discussed in AV ethics. Dietmar Hübner and Lucy White (2018) have made some progress in drawing a distinction between parties who are *involved* in the collision, and parties who are *uninvolved*. I take it that Hübner and White are echoing Thomson (1991: 298) here. Thomson's view is that the AV ought not swerve into Motorcyclist to avoid the threat to Passenger. According to Thomson, X is morally permitted to kill Y in self-defence only if, and because, Y would otherwise violate X's sufficiently important rights. Motorcyclist is what Thomson calls a *bystander*, which Thomson understands as a person who is not causally involved in the threat to Passenger's life.[47] Because Motorcyclist is a bystander, it is not the case that Motorcyclist will violate Passenger's sufficiently

[47] Strictly, this is a simplification. Thomson (1991: 298-9) allows bystanders to have a minimal causal role in the threat: 'First, if Y is in no way causally involved in the situation that consists in X's being at risk of death, then Y is clearly a bystander to it. And second, if Y is causally involved in it, and not minimally so – as, for example, when it is Y himself or herself who is about to kill X – then Y is clearly not a bystander to it.'

important rights unless the AV swerves and kills Motorcyclist. Hence, Passenger is not permitted to kill Motorcyclist in self-defence, and neither is the AV.

I am unconvinced by Thomson's view. There are, I think, two problems. First, it seems false that it is never morally permissible to kill bystanders. McMahan (2009: 172-3) illustrates this point with cases like this:

> *Evil Motorcyclist:* There is a massive truck headed towards the AV. Passenger will be killed in a collision with the truck. The only way to avoid the collision is for the AV to swerve into Evil Motorcyclist, who would be killed by the impact. Evil Motorcyclist has clocked the situation. He has long hated Passenger. He realises that this is a perfect opportunity to bring about Passenger's death. Though Evil Motorcyclist could speed up to avoid blocking the AV's escape route, he decides to stay in place to ensure Passenger's untimely death.

McMahan's insight is as follows. Evil Motorcyclist has exactly the same causal role as Motorcyclist. But whilst it seems impermissible to kill Motorcyclist, it does not seem impermissible to kill Evil Motorcyclist. So, Thomson is mistaken to hold that X is morally permitted to kill Y in self-defence only if, and because, Y would otherwise violate X's sufficiently important rights. McMahan's resolution to this problem is to distinguish between *culpable bystanders* and *innocent bystanders*. Whilst killing innocent bystanders is wrong, killing culpable bystanders is permitted.

There is another, related, problem. Many people think that killing bystanders is morally wrong. Thomson believes that what *best explains* the wrongness of killing bystanders is that it is false that the bystander would otherwise violate the victim's sufficiently important rights. Mike Otsuka (1994) disagrees. According to Otsuka, the wrongness of killing bystanders is better explained by the fact that bystanders are not morally responsible for the threat posed to the victim. Otsuka thinks that X is permitted to kill Y in self-defence only if, and because, Y is morally responsible for the threat posed to X's life. This explanation has a lot going for it. In particular, it explains why it is permissible to kill McMahan's culpable bystanders, as these people are in part morally responsible for the threat to the victim's life. Relatedly, it also explains the intuitive verdict in cases where the bystander is the indirect cause of the threat to the victim's life. For example, if Evil Motorcyclist had earlier programmed the massive truck to kill Passenger, and was now riding alongside Passenger's AV to watch Passenger be killed by the massive truck. Presumably, in this case it would be morally permissible for Passenger to kill Evil Motorcyclist.

For these reasons, I believe that we ought to reject Thomson's view. We should also reject Hübner and White's distinction between *involved* and *uninvolved* parties.

Helen Frowe (2014) defends a view that I think is much closer to the truth. For Frowe, the first question we ought to ask when deciding whether it is permissible to kill a person in self-defence is as follows: 'Does the person's actions, movement, or presence, contribute to the threat to the victim's life; or have the person's actions, movement, or presence already contributed to the threat to the victim's life?' (Frowe 2014: 31). If the answer is No, then the person is either an *onlooker* or a *bystander*. The difference between onlookers and bystanders is that harm directed at onlookers will not avert the threat to victim. But harm directed at a bystander will avert the threat to the victim. To illustrate: an onlooker might be a pedestrian standing on a bridge observing the collision scenario. On Frowe's view, it is morally wrong to kill onlookers. These people have a right not to be killed, and there are no grounds to defeat the presumption against killing them. Bystanders are people the killing of whom would avert the threat. For example, there might be a large pedestrian standing on the sidewalk who could be pushed in front of the massive truck to bring it to a stop before it hit the AV. Frowe also thinks that it is wrong to kill bystanders. The only reason to kill bystanders is that it is expedient from the point of view of the victim. But expedience does not justify killing someone. In short, killing bystanders is *exploitative*, and exploitative killings are wrong.

According to Frowe, Motorcyclist is neither an onlooker nor a bystander. This is because Motorcyclist's presence contributes to the threat to Passenger. After all, had Motorcyclist not been there, the AV could easily avoid the massive truck. The next question to ask is: 'Will the person kill the victim if the victim or a third-party does not kill them first?'. If Yes, then the person is a *direct lethal threat*. If No, then the person is an *indirect lethal threat* (Frowe 2014: 43-5). Motorcyclist is an indirect lethal threat. The reason why is that Motorcyclist will not kill Passenger if he is not killed first by Passenger or a third-party. Frowe argues that it is permissible to kill indirect lethal threats only if, and because, they have availed themselves of a reasonable opportunity to avoid posing the threat. Whilst Motorcyclist counts as an *innocent indirect lethal threat*, Evil Motorcyclist is a *responsible indirect lethal threat*. Thus on Frowe's view it is permissible for the AV to kill Evil Motorcyclist, but it is impermissible to kill Motorcyclist. This nicely captures the two criticisms that McMahan and Otsuka raised against Thomson's view, insofar as the AV is morally permitted to kill the motorcyclist only if he is in some way morally responsible.

I accept Frowe's distinction between innocent and responsible indirect threats. Thus I accept the *structure* of Frowe's view. Where I part company with Frowe is on the conditions under which a person counts as morally responsible for posing

an indirect threat. According to Frowe, a person counts as morally responsible if they fail to take a reasonable opportunity to avoid posing the threat. Consider,

> an agent counts as *intentionally failing* to take a reasonable opportunity to avoid posing a threat of unjust harm to Victim if (i) Victim has a right not to suffer the harm, (ii) the agent believes that a given course of action will endanger Victim, (iii) the agent believes that she has alternative courses of action available to her that are not unreasonably costly to her or other innocent people and that will not endanger Victim (or will not endanger him to the same degree), and (iv) she chooses not to take any of those alternative courses of action. Endangering Victim includes helping to bring about a harm to Victim or obstructing a valuable escape route of which Victim could avail himself (Frowe 2014: 73).

Frowe adds that 'it is unlikely that we will be able to establish a clear threshold at which the cost to an agent becomes unreasonable.' However, 'what is reasonable is sensitive to both the cost to the agent of taking the opportunity *and* the prospective harm to Victim that taking the opportunity avoids' (Frowe 2014: 77). The basic ideas here seem correct to me. But Frowe's conditions are problematic in cases that involve reasonable opportunities presented over time. Consider,

> *The Next Reasonable Opportunity:* There is a massive truck headed towards the AV. Passenger will be killed in a collision with the truck. The only way to avoid the collision is for the AV to swerve into Indecisive Motorcyclist. Indecisive Motorcyclist has clocked the situation. There are only two opportunities for Indecisive Motorcyclist to get out of the way. First, he could swerve into a ditch immediately. Because of his speed, Indecisive Motorcyclist knows that this will cause some broken bones. Second, Indecisive Motorcyclist can see a lay-by up ahead. He is reasonably certain that he can make it to the lay-by before the massive truck collides with the AV, and thus leave room for the AV to avoid the crash. He is also reasonably certain that he can avoid any injuries at all if he goes for the lay-by rather than the ditch. Thus Indecisive Motorcyclist decides to wait for the lay-by, and in doing so, endangers Passenger to a greater extent.

Frowe's view implies that it is morally permissible for the AV to kill Indecisive Motorcyclist as soon as he passes the ditch. Once Indecisive Motorcyclist passes the ditch, he has availed himself of an opportunity to pay a reasonable cost to avert the threat to Passenger. (If some broken bones seems too extreme for the cost to be reasonable, the reader is welcome to substitute in a lesser cost.) The trouble is that Passenger is now endangered to *a greater degree* than she was before. The AV must now swerve in the nick of time once Indecisive Motorcyclist moves into the

lay-by. Hence Indecisive Motorcyclist has passed up all the 'alternative courses of action that are not unreasonably costly [...] and that will not endanger Victim (or will not endanger [her] to the same degree).' He has chosen instead to take a later reasonable opportunity that he knows endangers Passenger to a greater degree.

Really, the problem is that Frowe thinks that the reasonableness of a cost can be determined in isolation. Frowe does not take into account the variation in the costs to Indecisive Motorcyclist across multiple opportunities that present through time, and how different costs to the Indecisive Motorcyclist map to different costs to Passenger. I think that it would be reasonable to require Indecisive Motorcyclist to drive into the ditch if that were the only option. But in light of the fact that there is another option that is much better from Indecisive Motorcyclist's point of view, and not significantly worse from Passenger's point of view, I do not think it reasonable to require that Indecisive Motorcyclist drives into the ditch.

We can now extract some rules for obstructer cases. The AV is permitted to kill a road-user who blocks an escape route only if that road-user has failed to take a reasonable opportunity to move. Which opportunities count as reasonable depend on the costs to the obstructer over the different opportunities available, and the costs to the AV's passenger. Obstructers are required to bear significant but non-lethal costs to avoid posing an indirect lethal threat to the passengers in the AV.

In practice, AVs will not know what other road-users believe. Presumably, AVs will also have limited computational resources to model the escape routes available to road-users who are blocking the AV's escape from a threat to its passenger. How does the deontological account address these uncertain cases? Because what is at stake is the potential killing of innocent persons, AVs ought to veer on the side of caution. First, if the placement of objects in the environment is such that there is no obvious means for the obstructer to remove themselves from the situation, then the AV should act conditional on the assumption that the obstructer cannot move. Here the AV's passenger ought to bear the cost, as it is reasonable to believe given the AV's evidence that the obstructer is an innocent indirect threat. Second, if there is reasonable doubt about whether the obstructer has an opportunity to remove themselves from the situation, the AV should act conditional on the assumption that the obstructer is an innocent indirect threat. Here the passenger should again bear the costs. Third, I think the only situation in which it is justifiable for AVs to kill obstructing road-users in practice is where the environment is such that there is ample opportunity for the obstructer to move. The AV morally ought to give the obstructer reasonable time in which to move, and do what it can to alert them of the threat. For example, sounding the horn. If, and only if, the obstructer has still

failed to move in these circumstances, then it is reasonable to believe given the AV's evidence that the obstructer is a responsible indirect threat; and it is permissible for the AV to kill the obstructer to avoid the independent threat to its passenger.

There are more complexities. The AV will be uncertain about the degree of harm that will result from colliding with an obstructing road-user. Though we cannot trade in certainties, there are certain heuristics that can guide us through. First, the harm that a vulnerable road-user such as a pedestrian or cyclist can be expected to sustain in a 30mph or greater collision is serious. Accordingly, the presumption against colliding with obstructers is strongest if they are vulnerable road-users. The presumption is somewhat weaker for motorcyclists, as these road-users wear protective equipment. The presumption is weaker still for obstructing vehicles; although the exact strength of the presumption will depend on the AV's rational expectation of the impact-velocity of the collision with the vehicle; and also the size of the vehicle in relation to the AV (the presumption is weakest for large trucks). What do these presumptions indicate? First, the stronger the presumption, the more confident the AV needs to be that the road-user can avoid obstructing the AV in its escape from the independent threat. Second, the stronger the presumption, the more time the AV needs to provide for the road-user to avoid posing the threat.

That concludes the treatment of obstructer cases. In short, the AV is morally permitted to kill an obstructer only if it is beyond reasonable doubt that they are deliberately obstructing the AV. Otherwise, the passenger should bear the cost. The AV should also do what it can to alert the obstructer before killing them; and its degree of aversion to killing the obstructer should correlate with the harm that it rationally expects the obstructer to receive conditional on its colliding with them.

5.4. Jaywalking Pedestrians

I imagine that many pedestrians who die on the road are jaywalkers, or people who cross the road in places where they should not. These people present interesting and difficult questions. When, if ever, are AVs morally permitted to kill or harm jaywalkers? How should AVs balance their passengers' safety against the safety of pedestrians who jaywalk? In this section, I shall address these questions.

I take it that pedestrians have a right not to be killed or harmed. Thus it is a presumptive moral wrong to kill pedestrians of any sort. That includes jaywalkers. The question is whether this presumption can ever be defeated. I shall argue Yes. In some rare circumstances it is permissible for AVs to kill or harm jaywalkers.

Sometimes bad arguments are instructive. We can learn from their mistakes. I shall take one such argument as my point of departure. This argument is unsound in multiple ways. But we can learn some important lessons from it. Consider,

> *Bad Argument:* Jaywalkers know that they are undertaking a risk of death or serious harm when they recklessly cross the road. In knowingly undertaking these risks, jaywalkers give tacit consent for AVs to kill or harm them. Because jaywalkers tacitly consent to these risks, they waive their rights against being killed or harmed by AVs. So, it is not wrong for AVs to kill or harm jaywalkers.

Obviously, this argument is problematic. There are, I think, three problems. The first is that the conclusion admits far too much. To see this, suppose that in

> *Unnecessary Killing:* Unlucky Jaywalker crosses the road where he should not have done. AV can stop to avoid the collision, leaving Unlucky Jaywalker with plenty of space to make it across the road safely. This would not endanger any other road-user, as there is nobody else for miles around. Nevertheless, AV accelerates into the collision and kills Unlucky Jaywalker instantly.

I shall take it as a given that the AV acts wrongly here. However, if it is true that jaywalkers tacitly consent to being killed or harmed, then Unlucky Jaywalker is not wronged here. This is because Unlucky Jaywalker would have waived his right not to be killed or harmed by the AV, and thus he would have no justified moral complaint against the AV killing or harming him. But this is clearly false. The AV at the very least acts wrongly in killing a jaywalker in a collision that could have been avoided at no cost to anyone else. So, there is good reason to reject the unqualified conclusion that jaywalkers tacitly consent to being killed or harmed by AVs when they cross the road. Instead we should appeal to the:

> *Better Argument:* Jaywalkers know that they are undertaking a risk of death or serious harm when they recklessly cross the road. In knowingly undertaking these risks, jaywalkers give tacit consent for AVs to kill or harm them in cases where this cannot be avoided without imposing serious costs on others. So, it is not wrong for AVs to kill or harm jaywalking pedestrians in cases where the only alternative course of action involves imposing significant costs on others.

This is much more plausible. But there is a second problem. The class of pedestrians to which this argument applies is too broad. Not all jaywalkers understand the risks of jaywalking. The jaywalker might have cognitive disabilities

that prevent them understanding the risks of jaywalking. The jaywalker may even be a young child. Clearly, it is false that *these* people knowingly undertake a risk of death or harm when jaywalking. The fact that some jaywalkers understand the risks of crossing the road and others do not needs to be accounted for. Consider,

> *Even Better Argument:* Some jaywalkers know that they are undertaking a risk of death or serious harm when they recklessly cross the road. In knowingly undertaking these risks, these jaywalkers give tacit consent for AVs to kill or harm them in cases where this cannot be avoided without imposing serious costs on others. So, it is not wrong for AVs to kill or harm jaywalking pedestrians who knowingly undertake risks of death or serious harm in cases where the only alternative course of action involves imposing significant costs on others.

This argument is an improvement. However, there is an immediate concern that the AV has no way of knowing whether the pedestrian does or does not understand the risks of jaywalking. I propose to bracket this concern for the time being. There is a far more serious problem with the Even Better Argument. The argument takes it for granted that knowingly undertaking a risk implies tacitly consenting to the manifestation of that risk. This is not true. Consider this counterexample:

> *Thomson's Mugging:* Suppose there are two ways in which I can get home from the station at the end of the day. The first is [...] safe, but is long. The second way is [...] unsafe, but is short. Nobody has ever been mugged while walking along Pleasant Way; people have from time to time been mugged on Unpleasant Way. Here I am at the station; I'm tired; I think 'The Hell, I'll chance it, I'll take the Unpleasant Way'. I then promptly get mugged (Thomson 1985: 139).

In this case, Thomson foresees the risk of being mugged. However, she has not consented to the mugging. Had Thomson consented to the mugging, the mugger would not have acted wrongly, as Thomson would have waived her right against being mugged. Yet the mugger did act wrongly. Therefore, knowingly undertaking a risk does not imply tacitly consenting to the manifestation of that risk.

There are two responses here. One is to deny Thomson's case judgement, and hold that Thomson did consent to the mugging after all. This is not a promising line of response. The other option is to turn the Even Better Argument inside out. Notice that Thomson's mugging case at most shows that knowingly undertaking a risk is consistent with not tacitly consenting to the manifestation of that risk. But it remains an open question whether it *can ever* be true that: (i) a jaywalker foresees the risk of death or harm; (ii) they tacitly consent to those risks; and (iii) their foresight of those risks provides a full or partial explanation of their tacit consent

to the manifestation of those risks. We know that X can explain Y even if X does not entail Y. My right not to be harmed explains why you ought not break my thumb. But there exist cases in which you ought to break my thumb despite my having that right. You might justifiably infringe my right if breaking my thumb is required to save several children from torture. Thus Thomson's case shows that we cannot take it for granted that jaywalkers tacitly consent to being killed or harmed by AVs in virtue of their foresight of certain risks. But this does not rule out building up to the conclusion of the Even Better Argument through a piecemeal approach. That is, an approach on which it is shown that jaywalkers in particular circumstances tacitly consent to harms in virtue of their foresight of the risks.

5.4.2. The Piecemeal Approach: Suicidal Pedestrians

I will begin the piecemeal approach with the simplest case in which it is plausible that a jaywalker gives tacit consent to being killed or harmed by the AV, where this tacit consent is explained at least in part by their foresight of the relevant risks. The moral considerations in this case are unfortunately quite complicated.

> *Suicide:* Suicidal Jaywalker decides that life is not worth living. She runs out onto the motorway in the hope of being killed by a vehicle. The AV has Passenger on board. The AV can either kill Suicidal Jaywalker, or swerve to avoid him, but in doing so kill Passenger by colliding with a lamppost.

Because there is a strong presumption against killing, I want to first ask whether Suicidal Pedestrian is in principle permitted to use lethal defensive force against Passenger should the AV not swerve. To make this easier, imagine that Suicidal Pedestrian has a ray-gun that he can use to vaporise the AV and Passenger.

Passenger might be counted as what Nozick called an *innocent shield of a threat.* These people are 'persons who themselves are nonthreats but who are so situated that they will be damaged by the only means available for stopping the threat.' The classic example that Nozick gives are 'Innocent persons strapped onto the front of the tanks of aggressors so that the tanks cannot be hit without also hitting them' (Nozick 1974: 35; see also Thomson 1990: 369–71; Walzer 1977: 172–5, esp. 174f).

What is the moral status of innocent shields of threats? You might think that the AV is the *real* threat to Suicidal Pedestrian, and that Passenger just happens to be inside the object that poses the threat. One argument for this view is that, had Passenger not been in the AV, the threat to Suicidal Pedestrian would still exist.

This argument, if plausible, suggests that Passenger is a *bystander* in Frowe's sense of the term, i.e. a person who does not threaten the victim, where the killing of this person would nevertheless avert the threat to the victim (c.f. Frowe 2014: 31).

However, as Gerald Lang (2007: 21) has argued, the counterfactual argument is not plausible. Lang contends that the counterfactual argument fails in cases that involve the object threatening the victim in its capacity as an object, i.e. by crushing them to death. According to Lang, Passenger and the AV form a 'composite object,' and it is this mereological fusion that presents a direct threat to Suicidal Pedestrian (c.f. Frowe 2008). We cannot say that the AV is the *real* threat, and Passenger just happens to be inside the threat, on the grounds that the threat to Suicidal Pedestrian would still exist even if Passenger were not inside the AV. This is because other parts of the composite object satisfy the same counterfactual condition. For example, the threat to Suicidal Pedestrian would still exist if we removed the AV's left-hand doors. This does not imply that the AV minus its left-hand doors is the *real* threat to Suicidal Pedestrian. So, on Lang's view, we cannot use the counterfactual condition to say that Passenger is not part of the threat.

Lang's argument suggests that the following claim is true: Passenger will kill Suicidal Pedestrian if Suicidal Pedestrian does not vaporise him first. That is, it is not the case that Passenger is just *inside* the object that will kill Suicidal Pedestrian. Passenger is a part of the object that presents the lethal threat. In Frowe's (2014) taxonomy, this renders Passenger a *direct lethal threat*. Obviously, it does not follow that Passenger is morally responsible for posing this threat. But Passenger might be considered *agentially responsible* for the threat, in the sense that the threat in some way traces back to Passenger's actions or presence (c.f. Frowe 2014: 74–5). Frowe's account holds that it is morally permissible to kill direct lethal threats. The reason for this is that killing direct lethal threats involves only *eliminative force*, as opposed to the *exploitative force* that is required when killing or harming a bystander (Frowe 2014: 64; c.f. Tadros 2011: 242–46). What is problematic about killing bystanders is that it is exploitative or expedient to kill them to save one's own life. In contrast, killing direct lethal threats – even if they are morally innocent – is not exploitative. Thus on Frowe's view, it is morally permissible to kill innocent shields of threats.

Frowe's view may be plausible in cases where the victim plays no part in bringing about the threat. But *Suicide* is different. Here Suicidal Pedestrian is the proximate and morally relevant cause of the AV's forced choice. My intuition in this case is that it is unfair for Passenger to bear the costs of Suicidal Pedestrian's decision to run onto the motorway. One plausible explanation for this intuition is

that Suicidal Pedestrian has given tacit consent to be killed by the AV. In support of this claim, we can point to the fact that Suicidal Pedestrian foresaw the risk of running onto the motorway and decided to run in the hope that those risks would manifest. If there is *any* example of a jaywalker tacitly consenting to being killed, presumably this is it. Accordingly, we could argue that Suicidal Pedestrian has, by giving tacit consent to be killed, waived her right against Passenger not to kill him. This would render it permissible for the AV to collide with Suicidal Pedestrian, and capture the intuition that Passenger is not morally required to bear the costs.

I am not entirely happy with this argument. The elephant in the room is whether Suicidal Pedestrian can consent to being killed by the AV. I mentioned in §5.2 that there are difficult questions about exactly what can and cannot be consented to. In the dispute over the moral permissibility of voluntary euthanasia, it is standard to add in certain caveats to the consent condition. For example, that dying is what is prudentially best for the person, and that the person's death will not be worse for others (McMahan 2002: 463). Accordingly, it is suspect that we should allow Suicidal Pedestrian's tacit consent to count as valid consent in the absence of *any* other conditions being met. But this is of course what the above argument requires.

Thankfully, there is a better option. There is a fine line between consent and forfeiture. Indeed, as Thomson (1990: 361) tells us, 'some waivings of rights might as well be called forfeitings.' The cases Thomson has in mind are ones in which the agent waives a right by 'letting it lie.' *Suicide* is one such case. Suicidal Pedestrian behaves with complete indifference towards her right not to be killed. Thus Suicidal Pedestrian might better be understood as having forfeited her right not to be killed. Her actions are such that she has no justified moral complaint against Passenger if the AV kills her and not Passenger. Thus we might argue that because Suicidal Pedestrian knowingly undertook the risk of forcing the AV to make this difficult choice, she forfeited her right not to be killed, and thus Suicidal Pedestrian is not permitted to defensively kill Passenger. In contrast, Passenger retains her right not to be killed by Suicidal Pedestrian. So, Passenger (and the AV) is morally permitted to kill Suicidal Pedestrian to avoid bearing the lethal costs of avoiding this act.

5.4.3. The Piecemeal Approach: Risk Takers

Can we go further? I think we can. The move from consent to forfeiture greatly improves the explanatory capabilities of our approach. There are cases in which,

though a pedestrian is not indifferent towards their right not to be killed, they are reckless with that right. Consider,

> *Interview:* Late Pedestrian is running late for an interview. He decides to take a shortcut by running across the motorway. Unfortunately, Late Pedestrian is slow. AV is travelling at 70mph. It will kill Late Pedestrian unless it swerves, in which case it will kill Passenger by crashing into a lamppost.

In this case, Late Pedestrian creates a dangerous situation. Passenger poses a direct lethal threat. But Passenger is not morally responsible for posing that threat. Instead, the fault lies with Late Pedestrian. Because Late Pedestrian's actions are the proximate and morally relevant cause of the AV's forced choice, I believe that Late Pedestrian is not permitted to use lethal defensive force against Passenger. For the same reasons, Passenger is not morally required to bear a lethal cost to avoid killing Late Pedestrian. Because Late Pedestrian is morally responsible for the forced choice, it is unfair for Passenger to have to bear the lethal cost.

However, there are limits to this argument. Because pedestrians are vulnerable road-users, and AV passengers are not, the harm that a jaywalker can expect to receive in a collision with the AV will often be greater than the harm that the AV passenger will expect to receive if the AV swerves. I have so far assumed that the passenger will die if the AV swerves. This will rarely be true. In practice, AV passengers are likely to be exposed to much smaller harms in evasive manoeuvres. Whilst I do not think that passengers are required to bear lethal costs to avoid killing jaywalkers, I do think that they are required to bear *some* costs. Consider,

> *Whiplash:* Impatient Pedestrian is walking through town. He crosses the road at an inappropriate place. Because he does not stop, look, and listen, AV is forced to either collide with Impatient Pedestrian at 40mph, most likely killing him; or swerve into a garden wall, causing Passenger minor whiplash.

I take it that Passenger ought to bear some cost to avoid killing Impatient Pedestrian, even though Impatient Pedestrian is morally responsible for the AV's choice. Passenger poses a direct lethal threat to Pedestrian. This is morally salient even if it is not the case that Passenger is morally responsible for the initial threat. Because Passenger poses a direct threat, there is some onus on Passenger to take steps to avoid posing the threat. Presumably, Impatient Pedestrian is permitted to use lethal defensive force against Passenger if Passenger fails to take a reasonable opportunity to avoid posing the direct lethal threat. What best explains this moral permission is that Passenger *becomes* morally responsible for posing the threat if he fails to take a reasonable opportunity to avoid posing the threat. Accordingly,

Passenger is not permitted to kill Impatient Pedestrian to avoid the whiplash; and the AV ought to swerve so as to avoid killing Impatient Pedestrian.

5.4.4. General Conclusions

AV passengers are not required to pay a lethal cost to avoid killing a jaywalking pedestrian. So, if the AV's choice is between killing a jaywalker and killing the passenger, the AV is permitted to kill the jaywalker. However, in practice, there is likely to be an asymmetry in the harms that the pedestrian or the passenger would suffer in a collision. Other things being equal, the harm that a pedestrian will sustain in a collision with the AV is far greater than the harm that the passenger would sustain in an evasive collision. Accordingly, provided the AV is travelling at roughly 30mph or greater, the AV should act on the assumption that the pedestrian will be killed or seriously harmed by the impact. The AV's passenger is required to undertake substantial costs to avoid killing or harming pedestrians, even if the pedestrian is morally responsible for bringing about the AV's forced decision. So, the AV ought to do everything it can to avoid colliding with the road user up to but not including imposing serious costs on its passenger (or other road-users).

5.5. Loss of Control Cases

I now consider cases in which the AV has diminished control over its speed and position. For example, cases in which the AV skids on ice; and cases in which the AV has a hardware or software malfunction. Cases of this latter sort include brake and steering failures. What does morality require in cases where the AV can lessen or eliminate the threat to its passenger by imposing a threat on another road-user?

5.5.1. McMahan, Kauppinen, and the Conscientious Driver

In a recent paper, Antti Kauppinen (forthcoming) discusses what morality requires in loss-of-control cases involving AVs. Kauppinen applies McMahan's theory of defensive killing to these situations. This view will act as the foil for my own view. Though I reject McMahan's view, I do not think that Kauppinen brought out its best qualities. I shall try to provide a charitable reconstruction of McMahan's view.

Consider,

> *The Conscientious Driver:* A person who always keeps her car well-maintained and always drives carefully and alertly decides to drive to the cinema. On the way, a freak event that she could not have anticipated occurs that causes her car to veer out of control and in the direction of a pedestrian (McMahan 2009: 165).

McMahan argues that the pedestrian is morally permitted to use lethal defensive force against the driver. This means that, if the pedestrian had a ray-gun, she would be morally permitted to vaporise the driver and the vehicle to eliminate the threat. McMahan's argument for the driver's liability to defensive force is complex. To be charitable to McMahan, it is helpful to pitch the argument in its proper context.

McMahan's big project is to reduce the morality of killing in war to the ordinary moral principles which govern killing in self-defence (McMahan 2009: 32-37; see also Frowe 2014: 123-4; Lazar 2009: 699-700). This approach to the ethics of war stands in contrast to the orthodox view that war is a *sui generis* moral context with its own rules for permissible killing (Walzer 1977; Zohar 1993). One of the many upshots of McMahan's approach is that it ties together the morality of going to war (*jus ad bellum*) with the morality of killing in war (*jus in bello*). McMahan challenges the traditional doctrine of the moral equality of combatants by arguing that the moral permissions of soldiers with respect to killing and harming others depend on their moral justification for going to war in the first place. Accordingly, the account of defensive killing that McMahan provides has evolved in response to his dialectical needs in the ethics of war. I will start with McMahan's original view on self-defence and explain how he revised this view to arrive at the somewhat counterintuitive view about the driver's liability in the conscientious driver case.

Early McMahan (1993) thought that an agent is liable to defensive force only if they pose a threat for which they are morally responsible. The rationale behind this account is that justice requires prioritising the lives of the innocent. As Lazar (2009: 708) puts it, early McMahan 'focused on distributing the impending, unavoidable harm to the person who was most at fault for it coming about.' So, if X deliberately poses an unjust threat to Y, then Y is permitted to kill X in self-defence; and the reason for this is that X is morally responsible for the threat that is posed to Y.

This account ran into difficulties. Middle McMahan (2002) acknowledges that in some cases soldiers on the unjust side lack moral responsibility for posing unjust threats to the soldiers on the just side. This is because some unjust soldiers have legitimate excuses which diminish or eliminate their moral responsibility. The excuses that McMahan has in mind include 'non-culpable ignorance, duress, and

diminished responsibility' (McMahan 2002: 401; see also McMahan 2009: 115–122). McMahan relaxed the requirement that unjust threats need to be *morally responsible* in order to be liable to defensive force. Instead he opted for the view that what is required for moral liability is *agential responsibility* for posing the threat. Roughly, this means that the threat traces back to the voluntary and minimally rational behaviour of the attacker (McMahan 2002: 411–421; see also Frowe 2014: 74–75). On this revised view, X is liable to lethal defensive force because he voluntarily forced Y into a situation in which Y must either kill or be killed by X. This view provides a plausible basis for a moral permission for just soldiers to attack unjust soldiers even if those soldiers have excuses and so lack moral responsibility.

Even this did not quite work. Late McMahan (2005, 2009) flags that the agency account is too restrictive in that it rules out the permissible killing of unjust soldiers that pose no direct threat to the just soldiers. These people might include, *inter alia*, reservists, soldiers in non-combat roles, and soldiers who are deliberately trying not to kill combatants on the other side (c.f. Lazar 2009: 709–711). In response to this problem, McMahan argued that people are liable to lethal defensive force if they knowingly behave in ways that create a *risk* that others will be forced to choose between being killed by them or killing them. Cue the conscientious driver. What McMahan hoped to show in the conscientious driver case is that a non-culpable threat can be liable to defensive force *simply* by foreseeing the risk that they might end up forcing someone to choose between being killed or killing them. Consider,

> Although [the conscientious driver] does not intend to harm anyone, she does know that her action carries a small risk of causing a great though unintended harm. Although her act is of a type that is generally objectively permissible, and although she has taken due care to avoid harming anyone, she has bad luck: the risk she knew her act carried has now, improbably and through no fault of her own, been realized. Because she knew of the small risk to others that her driving would impose, and because she nonetheless voluntarily chose to drive when there was no moral reason for her to do so – in short, because she knowingly imposed this risk for the sake of her own interests – she is morally liable to defensive action to prevent her from killing an innocent bystander (2009: 166).

This argument reflects a strikingly original contribution to the ethics of killing. What makes this argument so interesting is that McMahan in effect invents a novel justification for liability. McMahan (2005: 394) is explicit that the 'unjust threat that [the driver] poses is not the result of wrongful intent, recklessness, or negligence; it is the result of sheer bad luck.' These terms are borrowed from the

criminal law. To better understand the standard of liability that McMahan is employing, it is useful to compare McMahan's liability to these other standards.

In English and Welsh law, crimes committed with intent carry the greatest degree of criminal liability. Following *R v Woolin* (1999), the jury is permitted to infer that the defendant acted with intent if they are convinced that the harm done was *virtually certain* to follow from the defendant's act.[48] Reckless offences carry a lesser degree of criminal liability. In *R v Cunningham* (1957), it was held that the defendant acts recklessly if they act despite a foreseen risk of harm.[49] This definition was revised in *R v Caldwell* (1982) to remove the subjective element. Recklessness was thus defined as the defendant having not given thought to the possibility that their actions could result in harm or damage to property.[50] But the subjective standard was later re-introduced in *R v G* (2003). This case held that the defendant acts recklessly if they take a risk which, given the circumstances as they are known to the defendant, is unreasonable to take.[51] Last, criminal negligence reflects an even weaker standard of liability. Following *R v Bateman* (1925), criminal negligence entails that the defendant 'showed such a disregard for the life and safety of others as to amount to a crime … deserving of punishment.'[52]

That McMahan distinguishes his standard of liability from criminal recklessness and negligence is telling. McMahan assigns an ineliminable role to *luck*. There is no culpability on the driver's part whatsoever. The driver has merely taken a risk that she foresaw *could* result in the pedestrian being forced to either kill the driver or be killed by the driver. This is enough, on McMahan's view, to ground moral liability on the part of the driver should the risk manifest. The only standard of criminal liability that comes close to McMahan's view is *strict liability*. What makes an offence a strict liability offence is its lack of *mens rea* requirement: all that matters is that the defendant performed the act. Typically, strict liability offences have little or no stigma attached to them. For example, selling a lottery ticket to a minor.[53]

---

[48] [1999] 1 A.C. 82. Note: It is a common misconception that *Woolin* distinguished between two kinds of intent: direct intent and oblique intent. Here oblique intent is a lesser kind of intent marked by the defendant seeing the harm as a virtual certainty of their actions. It was clarified in *R v Matthews and Alleyne* [2003] Cr App R 30 that *Woolin* establishes an evidential standard that the jury can use to infer intent, and not a separate kind of intent.
[49] [1957] 2 QB 396.
[50] [1982] AC 341; see also *R v Lawrence* [1982] AC 510.
[51] [2003] UKHL 50.
[52] [1925] 19 Cr App R 8.
[53] See *Sweet v Parsley* [1970] AC 132. The scope of strict liability has expanded over time to cover some offences that do attract stigma. For example, in *R v G* [2008] UKHL 37, the House of Lords held that statutory rape of a child under 13 under s.5 of the Sexual Offences Act 2003 has a strict liability element. Lady Hale held that it was irrelevant whether the defendant believed or reasonably believed that the person he had sex with was over 13.

However, McMahan's notion of liability is stronger than strict liability, as he points to the fact that the conscientious driver 'knowingly imposed this risk for the sake of her own interests' (2009: 166). But this standard falls short of negligence.

I hope to have made clear what McMahan's view is, and why he holds this view. I next argue against McMahan's view, and in favour of a less stringent account of the moral permissions of the pedestrian and the conscientious driver. Then I will apply this revised view to the context of AVs who lose control, and in doing so, can mitigate the risk to the passenger by imposing risks on other road-users.

5.5.2. Revising McMahan's View

The main problem I have with McMahan's view is that I am unconvinced that foresight of the *possibility* of a collision is sufficient for liability to lethal defensive force. Frowe has made a similar point. According to Frowe (2014: 82), 'When a risk of endangering someone is sufficiently small, one can reasonably believe that one is not going to endanger anyone.' What is problematic here is that McMahan's account is insensitive to the probability of a fluke collision occurring. This does not sit well with moderate subjectivism about morality, the view that an agent's actions ought to be judged on the assumption that their reasonable beliefs about the morally relevant facts are true (c.f. Gibbard 1990: 42-3). The thought is that a passenger's reasonable beliefs about what will happen in the course of driving are sensitive to the probability that different events will occur. Because fluke accidents are improbable, McMahan's account seems to rely on an implausible objective standard of permissibility that does not provide a compelling basis for liability.

I am sceptical that Frowe locates the problem in the right place. There are cases in which foresight of a possibility, the probability of which is close to zero, renders agents not just *agentially responsible* but *morally responsible* when the risk manifests. Consider an example from the law. Suppose that X and Y are two men who decide to have sex. X is HIV positive and Y is not. X knows that he is HIV positive but does not tell Y. Because X and Y do not have any condoms, they have unprotected sex. X knows that the odds of transmitting HIV to Y are about 1 in 70 if X is the inceptive partner, and about 1 in 900 if X is the receptive partner. So, X decides to be the receptive partner so as to minimise the risk of transmission. Suppose that, against the odds, X transmits HIV to Y. Here X would be criminally liable for

reckless grievous bodily harm.[54] In this case, X has a *moderate* degree of criminal liability (the degree of criminal liability is between criminal negligence and intent). I want to suggest that X is also morally responsible here, although to a lesser extent than X would be if he had intentionally transmitted HIV to Y. However, if Frowe is right in her criticism of McMahan, then X would not be morally responsible.

So what is McMahan getting right? I think the problem is that Frowe appeals to the standard probabilistic conception of risk. On this view, the *risk* of an event is the probability that the event will occur; and the risk of an event is greater if its probability is higher (Ebert, Smith and Durbach 2019: 2). But I think that often our moral judgements track a *modal* notion of risk (Pritchard 2015, 2016). That is, an account of risk according to which the riskiness of an event is determined by how easily that event could occur. Here I shall assume the standard Lewisian setup, i.e. the closeness of a possible world is determined by its similarity to the actual world, such that the less that would have to change about the actual world to get to the relevant possible world, the closer that possible world is (Lewis 1973, 1979). Then we can say that an *easy possibility* is a possibility such that not much would have to change about the actual world in order to bring about that possibility. What I think McMahan is getting at is that when the driver drives for morally optional reasons, she foresees that she *could easily* end up imposing a lethal threat on a pedestrian. In some cases, I think that foresight of this kind is sufficient for moral liability.

Suppose the conscientious driver decided to drive for morally optional reasons in a snow storm on icy roads. Here it could easily be the case that the driver ends up imposing a lethal threat on a pedestrian. Though this is *unlikely*, not much would have to change about the world to make it happen. What would be required is for the car to skid and for a pedestrian to be in the wrong place at the wrong time. If the driver ended up in a skid, headed towards a pedestrian, I think the pedestrian would be permitted to use lethal defensive force against the driver. She could use the proverbial ray-gun to vaporise the car and the driver. I think that what best explains my case judgement is that the driver is morally responsible for the threat. The driver was reckless. She foresaw that she *could easily* pose a lethal threat to a pedestrian given the weather conditions, and chose to drive anyway for morally

[54] In *R v Dica* [2004] EWCA Crim 1103, the Court of Appeal held that a person who knows that they have HIV, fails to disclose it, and then subsequently infects a person, commits reckless grievous bodily harm (GBH) under s.20 of the Offences Against the Person Act 1861 (c.f. *R v Konzani* [2005] EWCA Crim 706). Interestingly, George Mawhinney (2013: 204) flags that there is a discrepancy between the *mens rea* required for reckless GBH for HIV transmission and reckless GBH in general. In the HIV case, the defendant is required to *know* that they have HIV. Standardly reckless GBH requires the defendant to *foresee* the risk of causing *some harm* (see also *R v G* [2003] UKHL 50).

optional reasons. This is a lesser degree of moral responsibility than if the driver had intended to kill a pedestrian. But I think it is a sufficient degree of responsibility for the driver to be liable to defensive force should she find herself in a forced choice.

There is a fine line between recklessness and negligence. Whether the driver in the snow storm case is morally responsible because of her foresight of the easy possibility of posing a threat, or because she showed a culpable disregard for the safety of others, is at best unclear. Nevertheless, what does seem to be clear is that the driver's moral responsibility for posing a threat can be established if, minimally, the driver drove for morally optional reasons with a patent disregard for the safety of other road-users. This will be true if it could easily be the case, given the road-conditions or the state of the vehicle, there are close-by possibilities in which the driver poses a lethal threat to a road-user. When this condition is met, I think that the driver is liable to lethal defensive force because they are morally responsible.

Can we take this further? I am sceptical. There is a case to be made for the view that *even if* the driver is not morally responsible for the threat, they are agentially responsible for a direct lethal threat. The driver's body is part of the mereological fusion that threatens the pedestrian. Because killing direct lethal threats requires only eliminative force, and not exploitative force, the pedestrian is permitted to kill the driver in self-defence (c.f. Frowe 2014: 74–5). However, as Lazar (2009: 714–28) has pointed out, determining who is agentially responsible for a threat is more complicated than McMahan, Frowe and others suggest. The problem is that their cases are too idealised. It seems implausible to suggest that the pedestrian has no agential responsibility for the dangerous situation. Really, the situation arose in light of the pedestrian's decision to go outside and the driver's decision to drive.

McMahan and Frowe seek to find small differences in the liabilities of the affected parties so as to justify inflicting harm on some but not on others. But my considered judgement is that in loss-of-control cases where the driver has not recklessly or negligently imposed a threat, there are no plausible liability grounds to inflict the harm on one party and not the other. The situation is an accident. If the brakes unexpectedly fail, the driver should take any reasonable opportunity to avoid killing a pedestrian. If they do not, then they will become morally responsible for the threat, and will thus be liable to lethal defensive force. But unless there are overriding lesser-evil considerations, e.g., if the car is headed towards a crowd of pedestrians, I do not think that the driver is morally required to bear lethal costs to avoid killing pedestrians in a collision for which they are not morally responsible.

What does all this imply about AVs in loss-of-control cases? First, insofar as it is possible, the AV ought to take into account facts about the road-conditions and its state of repair. If the AV has been taken out in adverse weather conditions, and these conditions are sufficiently serious as to amount to recklessness or negligence on the part of the passenger should a fluke collision arise, then the AV is morally required to impose the costs on the passenger and not on other-road users (to the extent that this is possible). Similarly, if the AV has certain faults which render likely an accident due to software or hardware malfunction, and it has notified the passenger about these faults, then the AV is required to impose the costs on its passenger in any collision that should arise because of these faults. Because having a positive moral reason to drive might counter-balance driving in adverse weather conditions, there are plausible grounds to exempt emergency vehicles from this requirement as it relates to weather conditions. In these cases, the AV should behave as it would if the passengers had not acted recklessly or negligently.

This takes us to the second point. If the AV loses control in a fluke collision *in general*, then neither the passenger nor the other road-users are liable to defensive force. The AV is required to impose substantial but non-lethal costs on its passenger to avoid killing other road-users. But the only situation in which the AV's passenger is required to bear a lethal cost is where there are strong lesser-evil grounds to avoid killing road-users. Because in general there is an asymmetry in the harms that a pedestrian will incur in a collision compared to the AV's passenger, the lesser evil will often be to inflict a serious harm on the passenger to avoid killing a pedestrian. The lesser evil condition will almost certainly be satisfied if the AV can avoid killing multiple pedestrians by imposing a lethal cost on its passenger.

5.6. In Defence of Deontology

I will do two things. The first is to tie together the points I have made into some general principles for the regulation of AV behaviour. The second is to spell out the abductive virtues which render my view more plausible than its competitors.

1. Road-users have rights against AV passengers not to be killed or harmed. This creates a strong presumption against the AV killing or harming road-users. This presumption is almost never defeated.
2. If there is a lethal threat to the AV's passengers, and there is some road-user blocking the AV's escape from that threat, then the AV is permitted to

kill that road-user only if it is beyond reasonable doubt that the road-user has failed to take a reasonable opportunity to move out of the way.

3. The AV's passenger is not required to bear a lethal cost to avoid killing jaywalking pedestrians. The AV is permitted to kill a jaywalker if the only other option is killing its passenger. But the AV's passenger is required to undertake substantial costs to avoid killing jaywalking pedestrians.
4. In general if the AV has partial control, it should opt for the lesser evil, and minimise aggregate expected harm over all affected parties. But if the AV's passenger is reckless or negligent, i.e. if they have taken the AV out in adverse weather conditions or if they have ignored fault warnings, then the passenger morally ought to bear the costs of a loss-of-control collision.

I shall now turn to the abductive virtues. The two virtues that I think are most important in this context are: (i) that our account of the morality of AV behaviour explains our considered judgements in the cases with which we are concerned; and (ii) that it coheres with our best theories of killing and harming in other domains.

First, the deontological account does a better job than other views at explaining our considered judgements. There is disagreement about how much this matters. For those in the tradition of W.D. Ross (1939: 1-2) and C.D. Broad (1930: 281-85), moral theories are intended as systematic accounts of common-sense morality. So, for these people, it matters greatly that moral theories explain our considered moral judgements. Others, like Henry Sidgwick (1907: 373), believe that the role of moral theory is not to describe our moral beliefs but to tell us which moral beliefs to have. But though Sidgwick is open to revisionary moral theorising, he accepts that moral theories must explain the bulk of our common-sense moral judgements. Thus though there are disagreements about the extent to which this first abductive virtue matters, there is broad agreement that it does matter, and that is good enough.

There has been limited success on this front in the ethics of AVs. First, several people defend *impartial* theories about the morality of AVs. These theories hold that permuting the identities or moral positions of the affected parties in a collision makes no difference to the moral status of the AV's acts. These theories include Derek Leben's (2017) 'Rawlsian' view and the act utilitarian account outlined by Jeff Gurney (2016). What is problematic about impartial theories is that these theories are insensitive to considerations pertaining to fairness and responsibility. On reflection, most reasonable people believe that considerations about who is morally responsible for the collision, or what is fair given the causal roles of the parties, makes a difference to what the AV morally ought to do. Because impartial theories set aside these considerations, these theories at most explain our

considered judgements about cases in which *all things are equal* between the parties. But this is one idealisation too far given that these cases will almost never arise.

There are also *partial* theories. These theories assign additional significance to the moral claims or moral preferences of the AV's passenger. For example, Jason Millar's (2014) view that AVs ought to allocate harm in accordance with the moral commitments of their passengers (c.f. Contissa et al. 2017; Keeling et al. 2019). This account and its cognates are plausible only if there are limits to the freedom that passengers have to determine the degree of partiality exercised by AVs. Otherwise these views would permit AVs to assign lexical priority to the safety of passengers, and kill pedestrians so as to avoid imposing minor risks of death or harm on their passengers. This is absurd. The obvious question is *how much* partiality the AV is permitted to exercise. Given the moral gravity of what is at issue – namely, killing the innocent, I suspect that the most plausible limits to impose are precisely those set by our best theories of permissible killing. But even if some other theory were best, the fact that these theories need to be restrained by another theory illustrates that these theories do a poor job of capturing our considered moral judgements.

In contrast to these theories, I think that the deontological account captures the bulk of our considered judgements about the morality of AV decisions. First, there is a clear explanation for certain general constraints that regulate AV decisions in the course of normal driving. For example, if the AV is stuck in traffic, it is wrong to drive on the pavement and kill pedestrians so as to get out of the traffic jam. The deontological theory provides a clear explanation for this: pedestrians have a right against the AV's passenger not to be killed, and the AV's killing them would violate this right. These explanations are hard to come by in the ethics of AVs. Second, in collision scenarios I hope that the deontological view captures to a great extent our intuitive common sense moral judgements about what the AV ought to do. Because the emphasis is on *who* is liable for *what* given their contribution to the predicament, the deontological view is sensitive to considerations of responsibility and fairness. Though there is room for reasonable disagreement on the finer details, I think that on balance the deontological view more or less captures our considered judgements.

I now turn to the second abductive virtue. The deontological view coheres with our best theories of permissible killing in other domains. In fact, it coheres with our best moral theories *tout court*, at least with respect to punishment, self-defence, war, medicine, sex, sport, and so on. Coherence with our best theories about other domains is not on its own important. Our best theories might be terrible. But coherence of this sort is indicative of a more fundamental abductive quality:

unification.[55] The view that I have defended is not unificatory; it is one corner of a much larger body of theory. But the general deontological theory of which it is a part ties together and systematises our considered moral judgements across a broad class of cases using a handful of basic moral concepts: claim rights, liability justifications, and lesser-evil justifications. That is quite something. There are some parts of morality, such as population ethics, where the deontological framework performs less well than consequentialism. It is not the final theory. But I am certain that any subsequent moral theory that adjudicates on matters of permissible killing must incorporate at least the basic structure of the position that I have defended.

The flipside is that it counts against the rival theories that these theories do not cohere with our best theories of killing in other domains. Take the impartial views. The inadequacy of maximising act utilitarianism as a theory of permissible killing is more or less clear. Recall Philippa Foot's (1967/2002) observation that whilst it seems permissible to kill one to save five in the classic trolley dilemma, it seems impermissible for a doctor to harvest the organs of one to save five others. Because utilitarianism conflicts with moral judgements like these, it is widely understood to be a non-starter in the ethics of killing. For Leben's (2017) 'Rawlsian' account, the results of importing this view into the ethics of killing more broadly are striking. Suppose that X attacks Y. Imagine that Y is permitted to kill X in self-defence only if, and because, killing X maximises the minimum survival probability for X and Y impartially considered given the options available to Y. This is not in the least bit plausible. Nor does it have *any* resemblance to our best theories of killing.

Things are worse for the partial views. Take Millar's (2014) theory. I cannot countenance a theory of killing in self-defence of the form X is permitted to kill Y in self-defence if, and only if, X endorses a set of moral principles that permit X to kill Y in self-defence. Or consider the view that X is morally permitted to perform a sexual act on Y just in case X endorses a set of moral principles that permit X to perform that act. These positions are deeply immoral. What is so problematic about theories of this form is that they give undue weight to autonomy – understood in the naïve sense of the agent's capacity for deliberation and choice, and no weight to the moral claims of the affected parties (O'Neill 2003: 3-6). The only way to render a view of this sort coherent with our best theories of killing is to impose restrictions on the moral preferences of agents, so that these preferences are ideal and track the requirements of morality (c.f. Harsanyi 1977: 631-36). But as we have known since

[55] See Philip Kitcher (1993) for the classic treatment of unification in the natural sciences.

Plato (2002: 12), idealised preferences are explanatorily redundant, as what matters is the moral independent standards that the preferences must adhere to.

That these other theories have little or nothing in common with our best moral theories about killing in other domains should give us pause for thought. In effect, the arguments for these theories better be paradigm-shifting. Otherwise there is no obvious reason to depart from the standard deontological theories of killing. But as I have argued in Chapters 1 and 4, these arguments are far from paradigm shifting. Because the rights that each of us has not to be killed by others are uniquely morally important, I suggest great caution before adopting novel and untested theories.

6. The Risk-Imposition Problem[56]

On March 18th 2018, Elaine Herzberg was killed by an automated vehicle (AV) whilst crossing the road with her bicycle in Tempe, Arizona. The AV detected an object on the road six seconds prior to the collision through its radar and LIDAR sensors.[57] It classified Herzberg as an *unknown*, then as a *vehicle*, and then as a *bicycle*. 1.3 seconds prior to the collision the AV determined that emergency braking was required; but its emergency brakes were disabled and the driver failed to respond in time (NTSB 2018). The Tempe collision invites some moral questions. What is the moral significance of object classification in AV decisions? What is the morally right amount of caution for AVs to exercise when uncertain about the classification of proximate objects? These questions are important aspects of what I call the risk-imposition problem. In this chapter, I shall defend answers to these questions.

In §6.1, I argue that the moral status of the AV's acts is indexed to its justified predictions about the classification of proximate objects. In §6.2, I put forward an account of how AVs ought to moderate their speed in road-traffic situations that involve uncertainty about what kinds of objects are in the AV's environment. The account holds that the AV ought to moderate its speed so that it can safely negotiate modally close *what if* cases such as a pedestrian walking out into the road suddenly. In §6.3 through §6.5, I examine what this theory might look like in practice. I

[56] This chapter is an extended version of my paper 'Automated Vehicles and the Ethics of Classification' that has been accepted to Ryan Jenkins, Tomas Hribek and David Cerny (eds.) *Autonomous Vehicle Ethics: Beyond the Trolley Problem*. Oxford University Press.

[57] Radar and LIDAR sensors use electromagnetic radiation to map the AV's environment. In both cases, the AV omits signals, and then detects signals that are bounced back by objects. The AV's object detection algorithms can use the doppler effect to estimate the motion of the object. Radar uses radio waves. The radar image is low-resolution. But radar is invariant under reasonable permutations in environmental conditions, i.e. it works in bad weather. LIDAR uses lasers. It provides a much higher-resolution image compared to radar. But it is considerably more expensive. For discussions of radar and LIDAR in relation to object detection see Cho et al. (2014), Chavez-Garcia (2015), and Beltran et al. (2018). Though radar and LIDAR data can inform the AV's *classification* of an object, cameras are best for object classification, as they have the highest resolution images. The major downside to cameras is that they do not work well at night (c.f. Ambardekar et al. 2008).

represent a mundane road-traffic scenario in which the AV is uncertain about an object's classification as a Markov Decision Process (MDP). I then examine how the account of risk-imposition that I defend might influence our choice of model parameters. In particular, I focus on the AV's reward function; and on its probability estimates for re-classifying objects at some future point in time. In §6.6, I conclude.

6.1. The Moral Significance of Classification

Consider,

> *The Fake Bush Case:* The AV detects an object 20m ahead at the side of the road. The object is round, stationary, and covered in leaves. Because the object has these features, the AV classifies it as a bush, and continues to drive at speed. The object is in fact a pedestrian dressed in a bush costume. Because the AV does not slow down as it overtakes, the pedestrian is exposed to a serious risk of harm.

How, if at all, does the moral status of the AV's acts depend on its predictions about the classification of proximate objects? In this section, I consider three views, and try to motivate the view that I believe is most plausible.

According to the o*bjective view*, the AV acted wrongly in failing to slow down. This is because, on this view, the moral status of the AV's acts depends on the morally relevant facts and not on the AV's internal representations of those facts.[58] The object was a pedestrian. AVs morally ought to slow down for pedestrians. The AV did not. So, the AV acted wrongly. Perhaps the agent responsible for the AV's acts might be *excused* given that the AV made a *reasonable* classification error. But the AV's failure to slow down was wrong even if the responsible party is blameless.

I do not accept the objective view. The objective view holds that the moral status of the AV's acts may differ across cases that are identical from the AV's perspective (c.f. Frowe 2010: 249-53; Gibbard 1990: 42-3). The AV ought to slow down in the Fake Bush Case, but not in a Real Bush Case in which the apparent bush is a bush. Thus it is impossible to *know* what the AV ought to do given the AV's information;

[58] The objective view standardly holds that an agent's moral obligations do not depend on their beliefs or justified beliefs about the morally relevant facts. Proponents of objectivism include Feldman (1988), Graham (2010), Moore (1903), Ross (1930), and Thomson (1986, 1990). I shall use the neutral term *internal representations* as opposed to *beliefs* so as to avoid attributing folk-psychological states to AVs. Though I think we could have some success predicting and explaining AV behaviour through the attribution of folk-psychological states, this approach is not terribly helpful for the ethical design of AVs (c.f. Dennett 1989). This is because it overlooks the details of how AVs represent features of their environments, and in doing so overlooks what I consider to be morally interesting considerations.

and the objective view implies that AVs must deliberate in accordance with what can be *excused* and not with what they *morally ought* to do.[59] Because our principal concern is the ethical design of AVs, I am unconvinced that there is much point invoking moral concepts and principles that cannot guide the ethical design of AVs (Goodall 2016; Himmelreich 2018; Keeling et al. 2019; Nyholm and Smids 2016).

What is needed is a subjective criterion of rightness.[60] That is, an account of permissible conduct for AVs that is indexed to the AV's internal representations of the morally relevant facts or to its evidence about those facts. According to *radical subjectivism*, the AV did not act wrongly in the Fake Bush Case. The rightness of the AV's acts should be judged conditional on the AV's internal representations of the morally relevant facts. Because the AV classified the pedestrian as a bush, and there is no moral reason to slow down for bushes, the AV did not act wrongly. According to *moderate subjectivism*, the rightness of the AV's acts ought to be judged conditional on the AV's epistemically justified or reasonable predictions about the morally relevant facts. Because the AV's evidence suggested that the object was a bush, and there is no reason to slow down for bushes, the AV did not act wrongly.

Of these two subjective views, I find the moderate view most plausible. Had the object in the Fake Bush Case possessed all the relevant features of a pedestrian, and had the AV nevertheless classified it as a bush, it would not have been permissible for the AV to continue driving at speed. What underpins this judgement is the intuition that the designers of the AV should have done better in training the AV's classifier algorithm to reliably classify objects which AVs can routinely be expected to encounter. Radical subjectivism fails to register the idea that some epistemic standard needs to be met in order for classification errors to render the AV's action morally permissible. For these reasons, I shall take moderate subjectivism as the point of departure in developing an account of the ethics of object classification.[61]

[59] See Mason (2012) for a discussion of this objection.

[60] For some good treatments of subjectivism see Frowe (2010), Gibbard (1990), Jackson (1991), Kagan (1989), Prichard (1932), Ross (1939), Scanlon (2008).

[61] I should flag that the views I have described are not necessarily incompatible. One view holds that there are multiple senses of the term 'ought.' The difference senses of 'ought' are the *objective ought*, the *belief-relative ought*, and the *evidence-relative ought* (c.f. Parfit 2011a: 150-64). What I think is true, minimally, is that the term 'ought' is used in each of these three senses; and that the correct semantics for the term 'ought' needs to take this into account. It is a deeper question whether the different senses of 'ought' correspond to morally interesting concepts; and in turn, whether one sense of 'ought' is fundamental. On pragmatic grounds I think the most appropriate standard of rightness here is moderate subjectivism.

6.2. Moderate Subjectivism for Automated Vehicles

In this section, I shall develop a moderate subjectivist view for object classification as it features in AV decisions. I shall focus on what is morally required in mundane road-traffic situations as opposed to collisions (c.f. Himmelreich 2018). I take it that the central moral question in these cases is balancing the AV's prudential goal of getting from A to B in reasonable time against the risk that faster driving imposes on proximate road-users.[62] The challenge is to explain how these competing values trade-off against each other to determine the appropriate speed for the AV given its uncertainty about the classification of proximate objects.

The value trade-off that we are dealing with is not new. Consider drunk driving. Many people think that drunk driving is morally wrong even if it harms nobody; and the best argument for this view is that drunk driving imposes undue risks of harm on road-users whereas sober driving does not (c.f. Husak 1994; Steinbock 1985; Oberdiek 2009, 2012). This argument takes it for granted that it is morally permissible to impose *some* risk on road-users in order to get to one's destination in good time; otherwise sober driving would be impermissible. But the argument also assumes that there is a limit to the amount of risk that drivers are permitted to impose; otherwise drunk driving would be permissible. The same holds for reckless driving. Whilst conscientious drivers are morally permitted to drive at *reasonable* speeds to get to their destinations in good time, reckless drivers act wrongly in driving at speeds which impose undue risks of harm on proximate road-users.[63]

The standard account of risk holds that 'the *risk* of a proposition P is determined by the *probability* of P – the higher the probability, the higher the risk' (Ebert, Smith and Durbach 2019: 2). I do not think that this conception of risk is helpful for our purposes. I think that our intuitive judgements about permissible risk-imposition are better captured with a *modal* notion of risk (Pritchard 2015, 2016). According to this view, a driver poses an unjustifiable risk to road-users if, and only if, given the vehicle's speed and the driver's evidence about relevant features of the driving environment, it could easily be the case that a road-user is injured by the vehicle. What could and could not easily be the case is determined by a closeness ordering

[62] For an excellent treatment of this issue from a technical perspective, see Sarah Thornton's section 'Value Sensitive Design for Motion Planning' in Keeling et al. (2019: 52-54).
[63] There is a more fundamental dispute about what explains why some risks are permissible and others are not. For example, it might be the case that imposing risks of harm is itself a form of harm, or that imposing serious risks of harm violates people's rights, or undermines their autonomy, and so on (Finkelstein 2002; Holm 2016; McCarthy 1997; Oberdiek 2009, 2012; see also Hayenhjelm and Wolff 2012 for a good overview of the dispute).

over possible worlds which reflects their similarity, or how much one world would need to change in order to be identical to another world (Lewis 1973, 1979). The idea is that a proposition could easily be true at this world if little or nothing would need to change in the actual world in order for that proposition to be true.

To illustrate, sometimes things happen on the road which are improbable, but given relevant features of the driving environment *could easily* occur. For example, if I am driving in a residential area, a child might run out into the road. This is not likely to happen. But it *could easily* happen. Thus when I am driving in a residential area I must be prepared to stop suddenly should a child run out into the road. Drunk driving imposes undue risks insofar as drunk drivers lack the competence to control their vehicles safely in challenging circumstances that could easily arise. Sober drivers, in contrast, can safely perform evasive manoeuvres or bring their vehicles to a sudden stop in a much broader class of challenging circumstances that could easily arise on the road. Thus, on the view that I defend, safe driving consists in the driver being prepared to stop in good time, or perform safe evasive manoeuvres, given the sorts of problems that could easily arise given the driver's evidence.

The modal view of risk has several advantages over the probabilistic conception. First, as Doug Husak (1994) points out, there is no significant difference between the probability that a drunk driver will kill a road-user and the probability that a sober driver will kill a road-user. Though drunk drivers have a higher *relative risk* of killing road-users; the *absolute risk* of killing a road-user for drunk and sober drivers is close to zero. Because absolute risk is what matters in this context, it is unclear that the probabilistic notion of risk explains our judgement that drunk driving imposes an undue risk whereas sober driving does not.[64] Second, when we reflect on the intuition that drunk or reckless drivers behave wrongly, what seems to best explain the intuition is that the driver shows a disregard for certain *what if* cases. For example, to someone who drives recklessly past a school, we might ask: 'What if a child had run out into the road?'. The reckless driver behaves wrongly

[64] You might think that there are some circumstances in which people act wrongly simply by increasing a relative risk. For example, suppose a parent from the UK travelled to Kenya with their child and failed to vaccinate the child against Yellow Fever. Yellow Fever is rare. So intuitively it is not true that the child could easily contract Yellow Fever. Furthermore, the rarity of Yellow Fever suggests that the parent's failure to vaccinate the child greatly increases the *relative risk* of the child contracting Yellow Fever. But the absolute risk does not change significantly. Thus it might be argued *contra* Husak (1994) that drunk drivers act wrongly *in virtue of* increasing the relative risk of killing another road-user significantly. I am grateful to Richard Pettigrew for this point. In response, I agree that this is a problem case for the modal view. But the salient question for our purposes is which notion of risk best explains our judgements in cases involving drunk or reckless drivers. Based on what I have said above, I think that the modal notion of risk is what best explains our judgements.

insofar as her speed is inappropriate given that a child *could easily* run into the road; and should that possibility manifest, she would be unable to safely avoid an accident.

I want to suggest a similar picture for the ethics of object classification. What matters morally is that AVs are driving with appropriate caution given plausible *what if* scenarios that could easily arise given the AV's evidence. Thus if an AV is overtaking an object, and it is severely uncertain about what kind of thing that object is, the AV ought to slow down *just in case* the object is a cyclist or pedestrian. Insofar as AVs are morally permitted to impose risks of harm to road-users given their prudential goal of time-efficiency, the AV must drive with the required level of caution so that it can come to a stop in *what if* cases that could easily occur given the AV's evidence about relevant features of the driving environment. Accordingly, I suggest that we take the following subjective standard of rightness as our guide for the ethics of object classification as it applies to AVs: 'The AV acts wrongly if, given its evidence about the classification of proximate objects, the AV's speed is such that the AV could easily injure a pedestrian in a close-by *what if* case.'

This principle is a standard of rightness; it is not a decision-procedure (Brink 1986: 421; Keeling 2020: 301-3). What this means is that the principle provides an explanation of why acts performed by the AV are right or wrong, but it does not provide instructions for how AVs should deliberate. In the remainder of the chapter, I consider what moderate subjectivism implies about the ethical design of AVs in practice. I must flag that the discussion is intended as a rough practical illustration.

6.3. From Theory to Practice

I said that the AV acts wrongly if, given its evidence, it could easily injure a road-user in a close-by *what if* case. In this section, I shall do what I can to explain what this view implies about AV decision-making in mundane road-traffic scenarios. The approach I take is to present a model of AV decision-making, and explain how this moderate subjectivist view might influence our choice of model parameters. The model involves some amount of idealisation. I am not proposing to build a workable AV decision-making algorithm. But I aim for the model to be close enough to the algorithms used in practice for the core lessons to carry over into the practical case.

6.3.1. Probabilistic Classifiers

First, I assume that AVs use probabilistic classifier algorithms.[65] Roughly, these algorithms classify objects in two stages. The first is to calculate the probabilities p_1 through p_K that the object belongs to each class c_1 through c_K, where each c_k is a class such as *pedestrian*, *cyclist*, *truck*, *car*, and so on. These probabilities are calculated in accordance with statistical methods that I propose to black box. The next step is to apply a decision-rule to the probability distribution over the different classes that the object might belong to. The standard rule is *maximum a posterior*, according to which the object should be classified as the most probable class. So, if there are three classes, pedestrian, cyclist, and truck, and for some object the AV assigns probabilities 60%, 30% and 10% respectively, then the object is classified as a pedestrian because this is the most likely class to which the object belongs. I add to this assumption that I shall consider simple cases in which the AV is uncertain about whether an object is a *pedestrian* or a *non-pedestrian*. The decision to focus on a binary classification problem simplifies the formalism considerably, and at little cost to the space of ethical considerations that the model invites us to examine.[66]

Second, I assume that the probabilities p_1 through p_K are calibrated. Often the probabilities that feature in probabilistic classification models are not reliable estimates of the true probability that the object belongs to each class.[67] The reason for this is that the maximum a posterior rule is fine for the purposes of classification if the probabilities rank the classes from most likely to least likely (Zadrozny and Elkan 2001). There is no gain from the perspective of accurate classification if the probability values used in the classifier provide more than a merely ordinal representation of the likelihood of the object belonging to each class. However, there exist techniques to extract calibrated probability estimates from a range of probabilistic and non-probabilistic classifiers. These include, *inter alia*, naïve Bayes classifiers, logistic regression models, decision trees, and support vector machines

[65] Formally, a classifier is a function from an n-dimensional vector space into a set of class labels. The vectors in the domain are called feature vectors. These vectors represent objects. Each component is a real-valued representation of a property belonging to the object. The classifier then matches each feature vector to an estimate of the object's class (Devroye, Györfi and Lugosi 1997: 1-7). Here I shall ignore the technical considerations that arise for building feature vectors for particular objects based on the AV's sensory input data.

[66] AVs have distinct pedestrian-detection algorithms that are independent from the algorithms used to locate vehicles and other objects (c.f. Thornton 2018: 118).

[67] Machine learning researchers do not tend to specify the interpretation of probability at issue when developing techniques for calibrated probability estimates. Here I shall assume that the *true* probability of the object being a pedestrian at a given time is the evidential probability of the proposition that the object is a pedestrian conditional on the evidence in the feature vector that is the AV's mathematical representation of the object's properties.

(Niculescu-Mizil and Caruana 2005; Platt 1999; Zadrozny and Elkan 2001, 2002). Thus in assuming that AVs can provide calibrated probabilities I hope that the discussion applies to a broad class of probabilistic and non-probabilistic classifiers.

6.3.2. The Markov Decision Process Model

I will discuss a simple mundane road-traffic situation that is easy to model. Suppose that the AV is driving along a country road. It detects an object up ahead, and this object might be classified as a pedestrian or a non-pedestrian. The AV can update its classification over time conditional on receiving new data about the object. The speed limit is 70kph/43.4mph. The task is to explain how the AV ought to moderate its speed as it drives past the object, taking into account the moderate subjectivist requirement that the AV should be able to come to a stop in close-by *what if* cases. The plausible *what if* case here is where the object is in fact a pedestrian, and the object steps out into the road, or for some other reason, enters the AV's path.

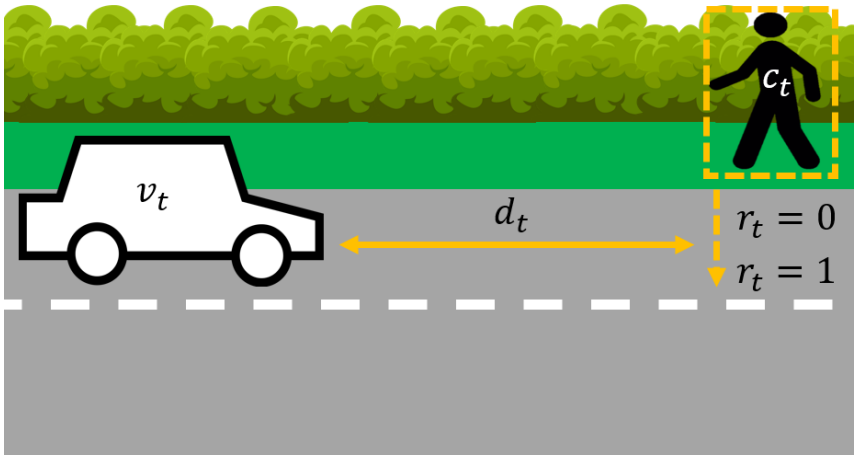

Figure 5: The Markov Decision Process Model

I will model this situation as a Markov Decision Process (MDP). I have chosen this model because AV engineers often use extensions of this model for AV motion planning (Bouton et al. 2018; Hubmann et al. 2018; Keeling et al. 2019; Thornton 2018; Ulbrich and Maurer 2013). There are limitations to the MDP model. There is a trade-off between, on the one hand, having a simple and transparent model that is reasonably accessible to those from a non-technical background; and having a model that best represents AV decision-making in practice. The natural model for representing AV decisions is the Partially Observable Markov Decision Process. But I think that on balance the MDP model is sufficient to make the core points,

even if this does require us to overlook certain real-world design considerations. The basic setup of the MDP model is illustrated in the diagram in **Figure 5**.

First, I assume that **time** is discrete. Events unfold in one second intervals. I refer to each time step as $t \in \{1,2,3,\ldots,T\}$. Second, at each time t, the AV is in a particular **state**. The AV's state is an epistemic state, in the sense that it describes the AV's predictions about all the relevant features of the situation.[68] I define a state as $s_t = (v_t, d_t, c_t, p_t, r_t)$. First, v_t is the AV's prediction of its speed at t, measured in metres per second; and d_t is the AV's prediction of its distance from the object at time t, measured in metres. Second, $c_t = 1$ if the object is classified as a pedestrian at time t, else $c_t = 0$; and the p_t term is the probability that the object is a pedestrian. These elements of the state are taken from the AV's classifier, and $c_t = 1$ if, and only if, $p_t \geq 0.5$. Last, $r_t = 1$ if the object is detected on the road at t, else $r_t = 0$. I assume to keep things simple that the unknown object is stationary in the horizontal direction: it can either remain in place or move onto the road.

Third, at each time t, the AV must choose an **act**, $a \in A$. I assume that the AV's acts are limited to changes in speed. The AV cannot change direction. The AV's act space contains options to accelerate or decelerate up to and including ± 3 metres per second squared. This includes the option for the AV to maintain constant speed. Fourth, when the AV performs an act a in state s, it **transitions** to a state s'. These transitions are probabilistic. The probabilities for state transitions depend only on the act chosen and the present state (c.f. Sucar 2015: 64). In formal terms,

$$T(s, a, s') = \Pr(s_{t+1} = s' \,|\, s_t = s, a_t = a)$$

I call $T: S \times A \times S \rightarrow [0,1]$ the **transition function**. I assume that transitions for d_t and v_t are deterministic.[69] First, $d_{t+1} = d_t - v_t$. This makes sense because the time-intervals are one second. Hence if the AV is travelling at v_t metres per

[68] Three Points: (1) The state-space as I have defined it is uncountably large. This is because p_t can take any value between 0 and 1. In practice, I would discretise p_t to make the state space finite. But it is fine for our purposes to imagine that it is uncountable. Note, however, that the value iteration algorithm described in the Appendix presupposes a finite state space. (2) It is non-standard to set up an MDP in this epistemic way. Normally, the states reflect the true states of the world rather than the true state of an agent's predictions. I am using the MDP in this way so as to avoid introducing an observation set, observation probabilities and a belief state, as is the case in Partially Observable MDPs. (3) It is possible to model POMDPs as MDPs with uncountable state spaces (i.e. a 'belief MDP'). But this is not what I am doing here. The AV's state is not a probability distribution over possible states of the system. The AV's state is a complete description of its relevant predictions at each time.

[69] These deterministic transitions are easily captured in the probabilistic transition function. The state s' that is the deterministic successor to s given act a is assigned probability 1 conditional on the AV performing a in s, and all other states are assigned probability 0.

second, and it starts at d_t metres from the object, then after one second it will be $d_t - v_t$ metres from the object.[70] Second, $v_{t+1} = v_t + a$. I assume that the AV's change in speed is instantaneous at the transition to the next time-step. I remain silent on the transition probabilities for c_t, p_t and r_t for the time being, as I will use these transition probabilities to explore certain ethical considerations later on.

The last ingredient is the **reward function**. The AV gets an immediate reward $R(s')$ for transitioning to state s' after performing act a in state s.[71] The AV's reward is a real-number. Hence $R: S \rightarrow \mathbb{R}$. The AV's aim is to maximise long-term expected discounted reward. The expected reward for performing act a in state s is the sum of the rewards received for transitioning to each state s' multiplied by the probability of transitioning to s' conditional on performing act a in state s.

$$\begin{aligned} \mathrm{ExpR}(a, s) &= \sum_{s' \in S} T(s, a, s')\, R(s') \\ &= \sum_{s' \in S} \Pr(s_{t+1} = s' | a_t = a, s_t = s)\, R(s') \end{aligned}$$

The solution to the MDP is a policy $\pi: S \rightarrow A$. This policy tells the AV to perform act $\pi(s) \in A$ in each state $s \in S$. Acting in accordance with π maximises the AV's long-term expected discounted reward. What it means for the reward to be discounted is that the AV assigns less significance to rewards further into the future. There are different methods for calculating π. In the **Appendix**, I explain one method. This is called *value iteration*. Here it is sufficient to say that the policy is computed as follows: the AV defines the value of a state $V(s)$ as the expected discounted reward for acting optimally in that state and in all subsequent states.

$$V(s) = \max_{a \in A} \sum_{s'} T(s, a, s') \left[R(s') + \gamma V(s')\right]$$

This is a recursive definition. We need to know the value of the subsequent states before we know the value of the present state. The value iteration method described in the **Appendix** solves the obvious problem that this presents. However, let us grant that the function $V(s)$ is well-defined. Then the policy π simply tells the AV to select the act in each state that it rationally expects to maximise expected reward provided it acts optimally in all states thereafter. In formal terms,

[70] The distances will be negative after passing the object.
[71] Often, the reward depends on both the state entered and the act performed to get to that state (c.f. Sucar 2015: 200). Here $R: A \times S \rightarrow \mathbb{R}$. We do not need this level of complexity.

$$\pi(s) = \underset{a \in A}{\operatorname{argmax}} \sum_{s'} T(s, a, s') \left[R(s') + \gamma V(s')\right]$$

The γ term in the two equations above is the AV's discount factor. The γ term is a number between 0 and 1. The closer γ is to 1, the more weight the AV assigns to rewards gained further into the future. I shall assume that $\gamma = 0.8$. This means that a reward of 10 in the present is afforded the same significance as a reward of 8 received at $t + 1$, 6.4 at $t + 2$, 5.1 at $t + 3$, and 1.7 at $t + 8$. The exact value of the discount factor is not important. What matters is that the AV takes into account expected rewards reasonably far into the future when deciding between actions.

6.4. Ethical Considerations for the Reward Function

I said earlier that the central moral question in mundane road-traffic scenarios is that of balancing the AV's prudential goal of getting from A to B in good time against the AV's moral goal of not imposing undue risks of harm on road-users. I suggested that the AV is permitted to pursue its prudential goal to the extent that, given its evidence, it does not risk harming road-users in close-by *what if* cases. The reward function is the part of the AV's decision-making algorithm in which this trade-off is most explicit (c.f. Keeling et al. 2019: 51-54; Thornton 2018: 73-74). I shall first present a plausible moderate subjectivist reward function for the AV; and then I shall discuss some of the ethical costs and benefits of this reward function.

First, I follow Thornton (2018) in pitching the AV's overall reward as the sum of multiple sub-reward functions that reflect different aims or values. Consider,

$$R(s') = R_{TIME}(s') + R_{SAFE}(s') + R_{STOP}(s')$$

Here R_{TIME} encourages the AV to drive at the speed limit. R_{SAFE} encourages the AV to drive at a speed such that it can safely handle modally close *what if* cases. In our case, this means a speed which enables the AV to stop in good time should the object unexpectedly enter the AV's path. Finally, R_{STOP} is the emergency stop. This sub-reward function encourages the AV to stop if the object enters the road.

I shall explain each part of the reward function in turn. Consider,

$$R_{TIME}(s') = -(19.4 - v_t)^2 \mathbf{1}(r_t = 0) \mathbf{1}(c_t = 0)$$

When the AV performs act a in state s, and transitions to state s', R_{TIME} gives a penalty equal to the square of the distance between its new speed and the speed-

limit. This in effect encourages the AV to drive at the speed-limit, i.e. to choose acts that will bring it closer to the speed-limit in the next step. The R_{TIME} reward is turned ON and OFF with the functions $\mathbf{1}(r_t = 0)$ and $\mathbf{1}(c_t = 0)$. First, $\mathbf{1}(r_t = 0) = 1$ if $r_t = 0$, and $\mathbf{1}(r_t = 0) = 0$ otherwise. Hence the time-efficiency reward is multiplied by 0 if the object is detected on the road in the new state. Second, $\mathbf{1}(c_t = 0) = 1$ if $c_t = 0$ and $\mathbf{1}(c_t = 0) = 0$ if $c_t = 1$. Hence the reward due to time-efficiency is multiplied by 0 if the object is classified as a pedestrian in the new state. These constraints on the time-efficiency reward capture the idea that the AV is permitted to pursue its prudential goal of time-efficient driving to the extent that, given its evidence, it could not easily harm a road-user. When the object is classified as a pedestrian, or when the object is detected on the road, the AV could easily harm a road-user by travelling in accordance with the time-efficiency goal. Hence it matters that the time-efficiency sub-reward function is sensitive to the AV's evidence about both the classification and location of the unknown object.

The R_{SAFE} reward reflects the AV's moral goal of driving in such a way that it could not easily injure a pedestrian in a close-by *what if* case. What counts as a safe speed cannot be determined in the absence of empirical data. Presumably, what is relevant here is the AV's stopping distance and the expected harm to a pedestrian conditional on collisions with different impact velocities. First, stopping distances for human drivers are divided into *thinking* and *braking* distances. I shall assume that AV thinking distances are close enough to instantaneous for us to discount them. The stopping distance for an average car at the 70kph speed-limit is about 30 metres. This reduces to 21m at 60kph; 15m at 50kph; and 9m at 40kph.[72]

Second, the expected harm to a pedestrian in a collision depends on multiple factors. These include, *inter alia*, the impact velocity of the collision, the size of the vehicle, and relevant features of the pedestrian such as age, weight, sex, height, and health condition (c.f. Anderson et al. 1997; Ashton 1982; Pasanen 1992; Davis 2001; Hannawald and Kauer 2004; Oh et al. 2008; Rosén and Sander 2009). I propose to keep things simple and focus on the probability of a fatality for an average adult pedestrian conditional on the impact velocity of the collision. I have extracted the following probabilities from Erik Rosén and Ulrich Sander's (2009: 539) logistic regression model.[73] If the pedestrian is hit at 70kph, the probability of a fatality is about 35%; at 60kph, the probability is 18%; at 50kph, 8%; and at 40kph, 4%.

[72] These stopping distances are calculated from the Random Science Tools (2017) stopping distance calculator. In line with the above, I have assumed a thinking distance of 0.

[73] The logistic regression model is $\mathrm{Pr}(\mathrm{Fatality}|v) = 1/1 + e\hat{}(6.9 - 0.090v)$.

The R_{SAFE} reward must strike a plausible balance between these factors. It is of course difficult to put numbers on things. The model is more precise than the morality that underpins it. But I would judge that a safe speed for the AV to travel at if it *reasonably predicted* that it was passing a pedestrian is in the region of 50kph. This in part reflects the fact that stopping distances for AVs are much shorter than the stopping distances for human drivers. 15m is a plausible stopping distance given the plausible *what if* case that the pedestrian might cross or fall into the road. The 50kph speed is also plausible in light of the fact that a fatal collision is less than 10% probable should the *what if* case manifest and the AV be unable to stop in time.

$$R_{SAFE}(s') = -(13.8 - v_t)^2 \mathbf{1}(r_t = 0)\mathbf{1}(c_t = 1)$$

Here 13.8m/s is equal to the safe speed of 50kph. The R_{SAFE} reward works in much the same way as the R_{TIME} reward. The AV is penalised for deviating from the desired speed of 13.8m/s. The penalty is equal to the square of the distance between the AV's speed in the state and the safe speed of 13.8m/s. The functions $\mathbf{1}(r_t = 0)$ and $\mathbf{1}(c_t = 1)$ are weights for the R_{SAFE} reward. When the AV performs act a in state s, and transitions to s', it receives a reward for safe driving only if the object is classified as a pedestrian and the AV has not detected it being on the road.

The final part of the reward function is R_{STOP}. This incentivises the AV to stop in order to avoid a collision with the object should the object enter the road. I do not think that this part of the reward function is morally interesting. Consider,

$$R_{STOP}(s') = -\left[\frac{v_t^2}{d_t + \varepsilon} + 50\mathbf{1}(d_t \leq 10)\right]\mathbf{1}(r_t = 1)$$

There are only so many ways to make a car stop. I have adapted this part of the reward function from Thornton's (2018: 73) research on designing AVs to navigate mundane road-traffic scenarios involving pedestrian crossings. The AV receives a negative reward for travelling at a high speed. The penalty increases as the distance between the AV and the object reduces. There is also a terminal penalty incurred if the distance between the AV and the object is less than or equal to 10 metres. The STOP reward thus encourages the AV to stop before it hits the object.[74] The $\mathbf{1}(r_t = 1)$ function ensures that R_{STOP} activates only if the object is detected on the road.

The reward function that I have described does a reasonable job at capturing the requirements of moderate subjectivism. The AV drives cautiously if its evidence suggests that the object is a pedestrian in a way that enables it to stop quickly if the

[74] Note that $\varepsilon > 0$ is a small buffer to avoid division by zero at $d_t = 0$.

object moves out into the road. But if the AV's evidence suggests that the object is not a pedestrian, then the AV drives in accordance with its time-efficiency goal. However, there is room for improvement. The main problem is that I have placed a great deal of trust in the AV's classifier algorithm. Recall that the AV's classifier predicts that the object is a pedestrian just in case it is more likely than not that the object is a pedestrian. Thus it might be true that the AV is 49% confident that the object is a pedestrian, and the object will be classified as a non-pedestrian. On the reward function described above, the AV assigns full weight to its time-efficiency goal even if the AV is 49% confident that the object is a pedestrian. This seems to overlook the fact that the moral costs of a false negative pedestrian misclassification are far greater than the moral costs of a false positive misclassification.[75] In short, it is far worse for the AV to drive fast when there is a pedestrian around than it is for the AV to reduce its speed somewhat when there are no pedestrians around. What is more, when the AV is severely uncertain about the object's classification, the object could easily be a pedestrian. Thus the AV plausibly acts wrongly in assigning full weight to its time-efficiency reward in these severe uncertainty cases.

There are plausible ways to modify the reward function in light of this point. One option is to *weight* the time-efficiency and safety rewards in proportion to the AV's estimate of the probability that the object is a pedestrian. Recall that the classifier is probabilistic. For each classification c_t, there exists a calibrated estimate of the probability p_t that the object is a pedestrian. The AV classifies the object as a pedestrian if, and only if, $p_t \geq 0.5$. Rather than turning the time-efficiency and safety rewards ON/OFF based on the classification prediction, we could instead weight the rewards by the probability that the object is a pedestrian in each state:

$$R'_{TIME}(s') = -(19.4 - v_t)^2(1 - p_t)$$

$$R'_{SAFE}(s') = -(13.8 - v_t)^2 p_t$$

This view assigns *greater* weight to time-efficiency the more confident the AV is that the object is not a pedestrian; and greater weight to safety the more confident the AV is that the object is a pedestrian. To illustrate, suppose that the AV is 49% certain that the object is a pedestrian, i.e. $p_t = 0.49$. Then when the AV performs act a in state s, and transitions to s', its reward $R(s')$ is determined as follows:

$$R'(s') = -[(19.4 - v_t)^2 0.51] - [(13.8 - v_t)^2 0.49] + R_{STOP}(s')$$

[75] See Haque (2012: 103) and Lazar (2018) for similar criticisms of simple threshold views in the ethics of war. For example, that a soldier is permitted to attack a person if it is more likely than not that they are an enemy combatant (c.f. McMahan 1994).

I shall ignore R_{STOP} for a moment. The weights on R_{TIME} and R_{SAFE} make it the case that the AV's reward is uniquely maximised just in case $v_t = 16.6$m/s:

$$\underset{v_t}{\text{argmax}}\{-[(19.4 - v_t)^2 0.51] - [(13.8 - v_t)^2 0.49] | 0 \leq v_t \leq 20\} = 16.6$$

What is salient here is that 16.6m/s is a *middling speed* between the AV's safe speed of 13.8m/s and its time-efficient speed of 19.4m/s. Thus the weights on each part of the AV's reward function ensure that the AV assigns additional significance to the possibility that the object is a pedestrian. This is true despite the fact that the object is classified as a non-pedestrian. Of course, when the AV is confident that the object is a pedestrian or a non-pedestrian, it acts primarily in accordance with the safety and time-efficiency rewards respectively. For example, suppose the AV is 90% certain that the object is a pedestrian, i.e. $p_t = 0.9$. In this case, the reward for entering state s' is uniquely maximised at 14.36m/s. In formal terms,

$$\underset{v_t}{\text{argmax}}\{-[(19.4 - v_t)^2 0.1] - [(13.8 - v_t)^2 0.9] | 0 \leq v_t \leq 20\} = 14.36$$

The reward-maximising speed of 14.36m/s is much closer to 13.8m/s, i.e. the ideal safe speed for the AV to travel at in the presence of a pedestrian. Obviously, this reward function has advantages over the original reward function. The reward is now sensitive to the AV's uncertainty about the object's classification. The AV is more cautious the more confident it is that the object is a pedestrian. Of course, there are other things that we could do to render the AV more risk-averse. Rather than weight the safety and time-efficiency rewards by the probabilities, we could instead weight these rewards by some function of the probabilities. Consider,

$$R''(s') = -[(19.4 - v_t)^2 (1 - \sqrt{p_t})] - [(13.8 - v_t)^2 \sqrt{p_t}] + R_{STOP}(s')$$

Here the R_{SAFE} reward is weighted by $\sqrt{p_t}$, and R_{TIME} by $1 - \sqrt{p_t}$. What is important here is that $\sqrt{p_t}$ is concave over the unit interval. This means that a line connecting any two points on the curve falls below the curve. Accordingly, $\sqrt{p_t}$ is a formal way to represent the idea that gains in the probability that the object is a pedestrian matter less and less for safe driving the higher p_t is in absolute terms. In other words, gains in p_t have diminishing marginal moral importance. Consider again the situation in which the AV is 0.49 confident that the object is a pedestrian. The original weighted reward $R'(s)$ assigned an optimal speed of 16.6m/s. But the risk-averse $R''(s)$ assigns a lower optimal speed of roughly 15.5m/s. Consider,

$$\underset{v_t}{\text{argmax}}\{-[(19.4 - v_t)^2(1 - \sqrt{0.49})] - [(13.8 - v_t)^2\sqrt{0.49}] | 0 \leq v_t \leq 20\}$$
$$\approx 15.5$$

Thus in using a concave weighting function, the AV is more risk-averse than on using the linear weights p_t and $1 - p_t$. Plausibly, this risk-averse set of weights better captures the idea that there is an asymmetry in the moral costs of false-positive and false-negative classification predictions. I shall end with this: I do not think that there is one decision-procedure that correctly captures the requirements of moderate subjectivism. That would be to suppose that morality is implausibly precise. Some individual judgement is required. I think it is true, minimally, that (i) the AV's classification predictions should not be taken at face value; and (ii) the AV morally ought to veer on the side of caution in its assignments of weights to time-efficiency and safety in situations where it is severely uncertain about whether or not an object is a pedestrian. I shall next turn to the AV's transition function.

6.5. Ethical Considerations for the Transition Function

Transition probabilities matter morally. The expected reward for performing an act a in state s is the reward received for transitioning to each state s' multiplied by the probability of transitioning to s' conditional on selecting a in s. Transition probabilities thus determine the weight given to morally salient possibilities such as the AV revising its classification prediction in the next time-step; and the object stepping out into the road at a subsequent time. In this section, I shall discuss what moderate subjectivism implies about the transition probabilities in our model.

Classification predictions might change. Recall the Tempe collision that I described at the start of the chapter. The AV classified Herzberg as an unknown, then as a vehicle, and then as a bicycle, in the seconds prior to the collision (NTSB 2018: 2). The suggestion here is that prior to the collision the AV was severely uncertain about its classification predictions. Plausibly, the probabilities for each class label were far from decisive; and the most likely class changed from moment to moment. In our model, the AV is concerned with two class labels: pedestrian and non-pedestrian. I want to suggest that if the AV is severely uncertain about the object's classification, then it could easily be the case that the object is a pedestrian. Thus the AV ought to exercise serious caution in these severely uncertain cases.

The AV's degree of caution can be expressed in the transition probabilities. To keep things simple, I shall assume that the AV is *severely uncertain* about the object's

classification just in case the probability p_t that the object is a pedestrian at time t is between 40% and 60%.[76] The exact numbers do not matter here. The moral question is this: When the AV is severely uncertain about its classification, what probability should it assign to its classification prediction being *the same* at $t + 1$?

I start with what we can call the *Risk-Neutral View*. This view says that when the AV is severely uncertain about whether or not the object is a pedestrian at time t, then it should assign equal or roughly equal probabilities to the propositions that it will classify the object as a pedestrian and a non-pedestrian at time $t + 1$. This position has a natural formal interpretation. We can say that the probabilities of the candidate values for p_{t+1} are determined by a random variable $P_{t+1} \sim \beta(5,5)$. $\beta(5,5)$ refers to a beta distribution with parameters $\alpha = 5$ and $\beta = 5$. Roughly, this is a bell curve over the unit interval with a midpoint of 0.5. This is shown in **Figure 6**. To be exact, **Figure 6** is a probability density function. The area under the curve for any interval $[a, b]$ is the probability that p_{t+1} falls into that interval. The total area under the curve is 1, i.e. p_{t+1} is certain to fall within $[0,1]$. Consider,

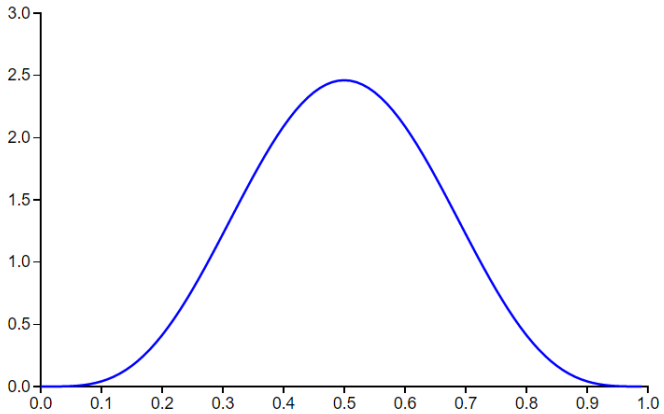

Figure 6: Probability Density Function for β-Distribution, α=5 and β=5

[76] The degree of uncertainty in the AV's probability distribution over the different classes for the object can be measured as the Shannon entropy of the distribution $(p_t, 1 - p_t)$. This is $\mathrm{H}(p_t, 1 - p_t) = -[p_t \ln(p_t) + (1 - p_t) \ln(1 - p_t)]$. The entropy ranges between 0 and $\ln(2)$. When the AV assigns equal probabilities to the object being a pedestrian and a non-pedestrian, the entropy is maximal and equal to $\ln(2)$. When the AV assigns a probability 1 or 0 to the object being a pedestrian, then entropy is minimal and equal to 0. My claim that the AV is severely uncertain when $0.4 \leq p_t \leq 0.6$ should be read as the claim that the entropy in the distribution is approximately $\ln(2)$. See MacKay (2003: 32) for discussion.

On the Risk-Neutral View, when the AV is in a state s_t, such that p_t is between 40% and 60%, the transition probabilities for p_t are as determined as follows:[77]

$$\Pr(a \leq p_{t+1} \leq b | 0.4 \geq p_t \geq 0.6) = \int_a^b \frac{1}{\mathrm{B}(5,5)} x^{5-1}(1-x)^{5-1}\, dx.$$

What this means is that if the AV is severely uncertain about its classification prediction *now*, then it is confident that it will continue to be severely uncertain in the next time step. The AV assigns *some* weight to the possibility that it will have a better idea of the object's classification in the next time step. But the AV is most confident that it will remain severely uncertain. The transition probabilities for the classification prediction c_{t+1} can be extracted from this distribution. I shall suppose that the probability that the object will be classified as a pedestrian in the next time-step is the expected value of the distribution, $5/(5+5) = 0.5$. Thus if the AV is severely uncertain about its present classification prediction c_t, it assigns equal probabilities to the object being classified as a pedestrian/non-pedestrian at $t+1$.

I do not think that the Risk-Neutral View captures the requirements of moderate subjectivism. The problem is that when the AV is severely uncertain about the object's classification, it *could easily* be the case that the object is a pedestrian. Really, I suspect that the AV ought to assign additional significance to the possibility that the object will be classified as a pedestrian in the next time-step. This is because there is a moral asymmetry between false positive and false negative classifications (c.f. Haque 2012: 103; Lazar 2018). When the AV is severely uncertain about the object's classification *now*, it should not behave as if it is *equally likely* that the object will be classified as a pedestrian or a non-pedestrian in the next time-step. Because the moral cost of misclassifying a pedestrian is far worse than the moral cost of misclassifying a non-pedestrian, the AV ought to exercise caution by weighting up the possibility that the object will later be classified as a pedestrian. Call this the *Risk-Averse View*. The AV ought to be risk-averse when it estimates how likely it is that the object will be classified as a pedestrian or non-pedestrian in the next time-step. The AV ought to assign additional weight to the morally salient possibility.

The parameters in the beta distribution can be changed to render the AV more risk-averse. **Figure 7** shows the probability density function for $P_{t+1} \sim \beta(8,4)$. Here the AV assigns additional confidence to the possibility that its future classification probability p_t will fall between 0.5 and 1. This in effect means that the AV weights up the possibility that it will classify the object as a pedestrian in the next time step.

[77] Here B is the beta function, i.e. $\mathrm{B}(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$. It acts as a normalisation constant.

Thus when the AV considers the anticipated rewards for actions several steps into the future, it will act on the assumption that it is reasonably likely that it will later classify the object as a pedestrian. The effect will be to weight up the R_{SAFE} reward over the R_{TIME} reward, rather than give equal significance to these rewards. This seems to me a plausible course of action from a moderate subjectivist point of view.

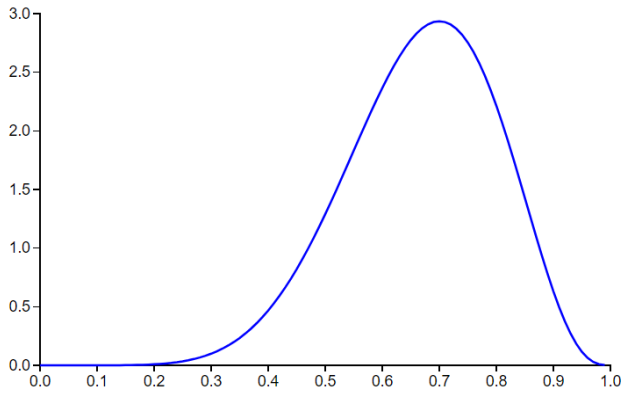

Figure 7: Probability Density Function for β-Distribution, α=8 and β=4

This distribution can then be used to extract transition probabilities for the AV's classification prediction c_{t+1} in the next time-step. The expected value of this distribution is $8/(8+4) = 0.66$. Thus the AV assigns a probability 0.66 to the object being classified as a pedestrian in the next time-step; and a probability 0.34 to the object being classified as a non-pedestrian in the next time-step. The effect here is that the AV will be more cautious than on the Risk-Neutral View, as the AV assigns greater significance to the possibility that the object will be classified as a pedestrian in later time-steps when performing its expected reward calculations.

This is a plausible interpretation of what moderate subjectivism implies about the transition probabilities in cases where the AV is severely uncertain about the object's classification. There is another morally interesting case to consider. How certain does the AV need to be that an object is a non-pedestrian before it is justifiable to set aside the possibility that it will be re-classified as a pedestrian?

I will say that the AV is confident that the object is not a pedestrian if p_t is at most 0.1. This means that the AV is at least 90% sure that the object is a non-pedestrian. I will make two claims. First, when the AV is confident at time t that the object is not a pedestrian, it is not unreasonable for the AV to assume that it will have similar confidence in the object not being a pedestrian when the object is reclassified in the next time-step. Second, when the AV is confident that the object

is a non-pedestrian, then it could not easily be the case given the AV's evidence that the object is a pedestrian. Because moderate subjectivism indexes the AV's moral permissions to its evidence, the AV is under no obligation to behave cautiously in these circumstances. From the AV's perspective, it could not easily be the case that it injures a road-user through its failure to slow down as it passes the object. Hence the AV is not morally required to reduce its speed under the stated conditions.

This might seem like an extreme view. This is especially true when we consider the moral asymmetry between the AV's passenger being delayed and the possibility that a pedestrian is exposed to an undue risk of death or serious harm. But I think that a degree of realism is required. We cannot have AVs slowing down for each post-box, tree, and lamppost, to which the AV assigns a small probability of being a pedestrian. There is a plausible criterion for determining when an AV is morally permitted to dismiss a risk as morally insignificant. This is where it could not easily be the case given the AV's evidence that its actions could injure a road-user.

How can we implement these claims? The simplest way to ensure that the AV assigns minimal weight to the possibility that the object will later be reclassified is through the transition probabilities. The AV is programmed to believe that when it is confident that the object is not a pedestrian, then it will continue to have that same level of confidence in subsequent time-steps. Of course, this can be overridden at any point if the AV revises its classification prediction unexpectedly. But so long as the AV acts under the apprehension that its estimate of the probability that the object is a pedestrian is unlikely to change, it will drive in a time-efficient fashion, and give little weight to the possibility that its classification will be revised.

To illustrate: One option is to model the transition probabilities for p_{t+1} as a random variable $P_{t+1} \sim \beta(0.5,5)$. Here P_{t+1} is such that the AV is almost certain to receive a low probability p_{t+1} from its classifier algorithm at time $t+1$. Consider,

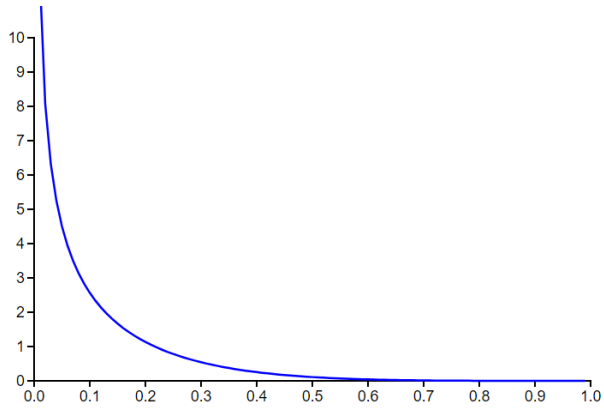

Figure 8: Probability Density Function for β-Distribution, α=0.5 and β=5

Here the AV assigns some weight to the possibility that it will become *less certain* that the object is a non-pedestrian. But it is confident that p_{t+1} will be close to 0. We can extract a transition probability for the AV's classification prediction c_{t+1} in the next time-step from this distribution. The expected value of the distribution is $0.5/(0.5 + 5) = 0.09$. This figure represents a good ballpark for the transition probabilities for c_t when the AV is confident that the object is not a pedestrian.

$$\Pr(c_t = 0 | 0 \geq p_t \geq 0.05) = 0.91$$

$$\Pr(c_{t+1} = 1 | 0 \geq p_t \geq 0.05) = 0.09$$

Obviously, these figures should be taken with a pinch of salt. The model is more precise than the morality that underpins it. But I hope that this provides a plausible moderate subjectivist answer to the problem of morally insignificant risks. Now it might be objected that I have not answered the question. The problem is how sure the AV needs to be that the object is a non-pedestrian before it is morally justified in setting aside the possibility that the object is a pedestrian. The truth is that it is difficult to put numbers on things. I am hesitant to go below 85% certainty that the object is not a pedestrian. Given the moral gravity of imposing risks of death and injury on innocent persons, AV designers ought to veer on the side of caution.

I shall end with a short discussion of the transition probabilities for r_t, i.e. the AV's prediction about whether or not the object is on the road. Given that the object is not on the road *now*, how much significance should the AV assign to its later being on the road? There are, I think, multiple factors to account for. First, suppose that the object is a pedestrian. Presumably, pedestrians are less likely to cross the road the closer the AV is. That is to say on common sense grounds that we can expect pedestrians to have a reasonable degree of road-safety training. In general, pedestrians are more likely to cross the road the further away the AV is in space.

Second, the probability of the pedestrian moving onto the road when the distance between the object and the AV is small is likely to be close to zero. But it *could easily* be the case that the pedestrian walks out into the road carelessly, or ends up on the road for some other reason, e.g. slipping on the grass bank and falling into the road. Hence there might be good reason to weight up the probability of the *what if* case, so that the AV assigns some significance to its having to perform an emergency stop at some point in the future. The effect of doing this would plausibly be a slight reduction in the AV's speed in the present. The following transition probability is intended to capture the points that I have made here. Consider,

$$\Pr(r_{t+1} = 1 | r_t = 0, 0 \geq d_t \geq 98, c_t = 1) = \frac{0.3}{0.3(100 - d_t)}$$

This means: 'The probability that the object will be detected on the road in the next time step conditional on the object being on the grass bank *now*, the distance between the AV and the object being between 0 and 98 metres, and the object being classified as a pedestrian, is $0.3/0.3(100 - d_t)$.' What does this entail? First, when the AV is far away from the object, it assigns a reasonably high probability to the object being on the road in the next time-step. For example, a probability 0.5 is assigned to the pedestrian crossing the road in the next time-step when the AV is 98 metres away from the object. This is an overestimate of the true probability. Presumably, pedestrians do not cross the road all that often. But in weighting up the probability, we force the AV to assign significance to the *what if* case in which the pedestrian does cross the road. This should cause the AV to reduce its speed in anticipation of the pedestrian crossing the road. However, the probability of the pedestrian entering the road diminishes rapidly as the AV approaches. At $d_t = 90$, the probability is 0.1. For $d_t = 70$, 0.03. For $d_t = 40$, 0.02. For $d_t = 10$, 0.01. In practice the probability of the object entering the road at $d_t = 10$ is much less than 0.01. But because this *could easily* happen, we should weight up the probability to encourage the AV to take into account this modally close *what if* scenario.

6.6. Conclusion

In this chapter, I discussed the risk-imposition problem in relation to cases in which the AV is uncertain about the classification of a proximate object. I defended the moderate subjectivist view that the AV's moral permissions in these cases are indexed to its justified or reasonable predictions about the classification of objects. I argued that the AV acts wrongly if, given its evidence, its speed and position make it the case that it could easily harm a road-user in a modally close *what if* case. Thus morality requires AVs to drive with sufficient caution, i.e. to moderate their speed, so as to enable them to safely negotiate these cases. I then discussed how this view might influence our choice of model parameters in a simple AV decision-algorithm. I showed how the AV's reward function can be tailored to reflect a plausible trade-off between its prudential goal of time-efficient driving and pedestrian safety; and how the AV's predictions about its future classification can be adjusted in order to register the moral asymmetry between false positive/false negative classifications.

6.7. Appendix: The Value Iteration Algorithm

This is a supplement to the Markov Decision Process model from Chapter 6. I will explain one method for determining a policy $\pi: S \rightarrow A$, such that choosing acts $\pi(s) \in A$ for each state $s \in S$ maximises the AV's expected discounted reward. This method is called *value iteration* (Bellman 1957; see also Sucar 2015: 202-203). The basic idea is to define the value $V(s)$ of each state $s \in S$ as the expected reward for acting optimally in state s, and then in all subsequent states thereafter. Hence the state value function $V(s)$ admits the following recursive definition. Consider,

$$V(s) = \max_{a \in A} \sum_{s\prime} T(s, a, s') \left[R(s') + \gamma V(s')\right]$$

The $\gamma \in [0,1]$ term is the AV's discount factor.

It is clear that $V(s)$ has a recursive definition. The value of state s depends on the values of the possible states s' that the AV might transition to. The policy π is such that it tells the AV to maximise expected discounted reward in the long-run.

$$\pi(s) = \operatorname*{argmax}_{a \in A} \sum_{s\prime} T(s, a, s') \left[R(s') + \gamma V(s')\right]$$

The AV is able to determine $\pi(s)$ for each state $s \in S$ using an algorithm that exploits a convergence result proved by Richard Bellman (1957). The algorithm has three stages: **initialisation**, **iteration, and termination**. The first stage at $k = 0$ is to set $V^{(0)}(s) \leftarrow 0$ for all states $s \in S$. The second stage is as follows:

`If` $\text{EndState}(s) = \text{True}$,

`then,`

$$V^{(k+1)}(s) \leftarrow 0$$

`else,`

$$V^{(k+1)}(s) \leftarrow \max_{a \in A} \sum_{s\prime} T(s, a, s') \left[R(s') + \gamma V^{(k)}(s')\right]$$

In plain English, if the state is an end-state, i.e. a state in which the MDP terminates, then its value at $k + 1$ is set to zero. There are no rewards to be gained in an end state. If the state is not an end-state, then its new value is computed using

the values of all the states as they are defined at stage k. Bellman proved that over time the values for each state s will converge. The iteration terminates once the state values have converged. In practice, there exists some small number ε such that the iterative process terminates if, and only if,

$$\forall s \in S: \left|V^{(k)}(s) - V^{(k-1)}(s)\right| \leq \varepsilon$$

The values $V^{(K)}$ for the terminal step $k = K$ are used for the AV's policy π so that the AV acts to maximise expected discounted reward.

7. Conclusion

I have defended a deontological picture of the ethics of automated vehicles (AVs). I argued that the AV is morally permitted to kill or harm a road-user if, and only if, its passengers are morally permitted to kill or harm that road-user in self-defence; and that AVs morally ought to drive with sufficient caution so that they can safely negotiate modally close *what if* cases such as young children running into the road. I argued that, on balance, this deontological account does a better job than its rivals at capturing our considered judgements in the cases with which we are concerned.

I hope also to have found a plausible methodological compromise between two approaches to the ethics of AVs that are often taken to stand in conflict. The first is a method that puts philosophical ethics centre stage; and the second is a method that puts the decisional algorithms used in AVs as the principal focus. On one hand, the deontological position that I have defended is a close cousin of our best theories of permissible killing in traditional ethical domains such as war and self-defence. On the other hand, in the final chapter, I tried to show how my view might influence our choice of model parameters in a stochastic control process that, though simple, is close in spirit to the algorithms that are used for AV decision-making in practice. I hope, minimally, to have shown that this hybrid approach to the ethics of artificial intelligence is more productive than disagreements about which method is best.

There are many unsolved problems. The problems that matter most relate to the research in Chapter 6. Moral philosophers must work alongside AV engineers to better understand AV decision-making algorithms. This is because it is unclear how different criteria for rightness are best reflected in these algorithms. What I hope to achieve in future is a deeper understanding of the statistical methods used in object classification, so that I can better appreciate the moral costs and benefits of the different model parameters that might be used in the design of AVs.

Bibliography

Alaieri, F., & Vellino, A. (2016). Ethical decision making in robots: Autonomy, trust and responsibility. *International Conference on Social Robotics* (pp. 159-168). Cham: Springer.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149-155.

Ambardekar, A., Nicolescu, M., & Bebis, G. (2008). Efficient vehicle tracking and classification for an automated traffic surveillance system. *Signal and Image Processing*, 1-6.

Anderson, R. W.-H. (1997). Vehicle travel speeds and the incidence of fatal pedestrian crashes. *Accident Analysis & Prevention, 29*(5), 667-674.

Aquinas, T. (2014). *The Summa Theologica.* New York, NY: Catholic Way Publishing.

Arnolds, E. B., & Garland, N. F. (1974). The defense of necessity in criminal law: The right to choose the lesser evil. *The Journal of Criminal Law and Criminology, 65*(3), 289-301.

Arrow, K. J. (1963). *Social Choice and Individual Values.* New York: John Wiley and Sons.

Ashford, E., & Mulgan, T. (2018). *Contractualism.* (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/sum2018/entries/contractualism/

Ashton, S. J. (1982). *SAE Technical Paper 801315: A preliminary assessment of the potential for pedestrian injury reduction through vehicle design.* Society for Automotive Engineers.

Austin, J. (1956/1961). A Plea for excuses. In J. Austin, J. Urmson, & G. Warnock (Eds.), *Philosophical Papers.* Oxford: Oxford University Press.

Bader, R. (2019). Person-Affecting Population Ethics (Unpublished Manuscript).

Bader, R. (forthcoming). Person-affecting utilitarianism. In G. Arrhenius, K. Bykvist, & T. Campbell (Eds.), *Oxford Handbook of Population Ethics.* Oxford: Oxford University Press.

Bales, R. (1971). Act-utilitarianism: account of right-making characteristics or decision-making procedure? *American Philosophical Quarterly*, *8*(3), 257-265.

Beauchamp, T. L., & Childress, J. F. (1994). *Principles of Biomedical Ethics* (Vol. 4). New York: Oxford University Press.

Beltran, J., Guindel, C., Moreno, F. M., Cruzado, D., Garcia, F., & De La Escalera, A. (2018). Birdnet: a 3d object detection framework from lidar information. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*.

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making. *Frontiers in Behavioral Neuroscience*, *12*, 31.

Berker, S. (2007). Particular Reasons. *Ethics*, *118*(1), 109-139.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy and Public Affairs*, *37*(4), 293-329.

Bjorndahl, A., London, A. J., & Zollman, K. J. (2017). Kantian decision making under uncertainty: Dignity, price, and consistency. *Philosophers' Imprint*, *17*(7), 1-22.

Boeglin, J. (2015). The costs of self-driving cars: reconciling freedom and privacy with tort liability in autonomous vehicle regulation. *Yale Journal of Law and Technology*, *17*, 171.

Bohlander, M. (2006). Of shipwrecked sailors, unborn children, conjoined twins and hijacked airplanes—taking human life and the defence of necessity. *The Journal of Criminal Law*, *70*(2), 147-161.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576.

Brandt, R. (1959). *Ethical Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Brandt, R. (1963). Towards a credible form of utilitarianism. In H. Castenada, & G. Nakhnikian (Eds.), *Morality and the Language of Conduct* (pp. 107-143). Detroit: Wayne State University Press.

Brennan, J. (2011). The right to a competent electorate. *The Philosophical Quarterly*, *61*(245), 700-724.

Brink, D. (1986). Utilitarian morality and the personal point of view. *The Journal of Philosophy*, 417-438.

Broad, C. D. (1930). *Five Types of Ethical Theory*. London: Routledge and Kegan Paul.

Broome, J. (1991). *Weighing Goods*. Oxford: Blackwell.

Broome, J. (1998). Review: Kamm on Fairness. *Philosophy and Phenomenological Research*, *58*(4), 955-961.

Broome, J. (2004a). Reasons. In R. J. Wallace, P. Pettit, S. Sheffler, & M. Smith (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (pp. 28-55). Oxford: Clarendon.

Broome, J. (2004b). *Weighing Lives*. Oxford: Oxford University Press.

Broome, J. (2013). *Rationality Through Reasoning*. Oxford: Wiley Blackwell.

Buchak, L. (2013). *Risk and Rationality*. Oxford: Oxford University Press.

Chavez-Garcia, R. O., & Aycard, O. (2016). Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, *17*(2), 525-534.

Cho, H., Seo, Y.-W., Kumar, B. V., & Rajkumar, R. R. (2014). A multi-sensor fusion system for moving object detection and tracking in urban driving environments. *2014 International Conference on Robotics and Automation (ICRA)* (pp. 1836-1843). IEEE.

Christie, G. C. (1999). The defense of necessity considered from the legal and moral points of view. *Duke Law Journal*, *48*(5), 975-1042.

Coco-Vila, I. (2018). Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. *Criminal Law and Philosophy*, *12*, 59-82.

Coleman, J. L. (1980). Efficiency, utility, and wealth maximization. *Hofstra Law Review*, *8*(3), 509-551.

Contissa, G., Lagioia, F., & Sartor, G. (2017). The ethical knob: Ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, 365-378.

Crama, Y., & Hammer, P. L. (2011). *Boolean Functions: Theory, Algorithms, and Applications*. Cambridge: Cambridge University Press.

Criminal Justice and Immigration Act. (2008).

Crisp, R. (1992). Utilitarianism and the life of virtue. *The Philosophical Quarterly, 42*(167), 139-160.

Crisp, R. (2003). Particularizing particularism. In B. Hooker, & M. O. Little (Eds.), *Moral Particularism* (pp. 23-47). Oxford: Oxford University Press.

Crisp, R. (2006). *Reasons and the Good*. Oxford: Clarendon Press.

Dancy, J. (1993). *Moral Reasons*. Oxford: Wiley Blackwell.

Dancy, J. (2003). The Particularist's Progress. In B. Hooker, & M. O. Little (Eds.), *Moral Particularism* (pp. 130-156). Oxford: Oxford University Press.

Dancy, J. (2004). *Ethics without Principles*. Oxford: Oxford University Press.

Davis, G. A. (2001). Relating severity of pedestrian injury to impact speed in vehicle-pedestrian crashes: Simple threshold model. *Transportation Research Record*, 108-113.

Davnall, R. (2020). Solving the single-vehicle self-driving car trolley problem using risk theory and vehicle dynamics. *Science and Engineering Ethics, 26*(1), 431-449.

Dennett, D. C. (1989). *The Intentional Stance*. Cambridge, MA: MIT Press.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.

Douma, F., & Palodichuk, S. A. (2012). Criminal liability issues created by autonomous vehicles. *Santa Clara Law Review, 52*, 1157.

Duffy, S. H., & Hopkins, J. P. (2013). Sit, stay, drive: The future of autonomous car liability. *SMU Science and Technology Law Review, 16*, 453.

Ebert, P. A., Smith, M., & Durbach, I. (2019). Varieties of risk. *Philosophy and Phenomenological Research*.

Eddington, A. (1920). *Space, Time and Gravitation - An Outline of the General Relativity Theory*. Cambridge: Cambridge University Press.

Evans, K., de Moura, N., Chauvier, S., Chatila, R., & Dogan, E. (ms). Ethical decision making for autonomous vehicles: The AVEthics Project.

Faden, R. R., & Beauchamp, T. L. (1986). *A History and Theory of Informed Consent*. New York: Oxford University Press.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., . . . König, P. (2019). Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and Engineering Ethics*, *25*(2), 399-418.

Feldman, F. (1988). Subjective and objective justification in ethics and epistemology. *The Monist*, 405-19.

Ferzan, K. K. (2005). Justifying self-defense. *Law and Philosophy*, *24*(6), 711-749.

Finkelstein, C. (2003). Is risk a harm? *University of Pennsylvania Law Review*, *151*, 963-1001.

Fleetwood, J. (2017). Public health, ethics, and autonomous vehicles. *American Journal of Public Health*, *107*(4), 532-537.

Fletcher, G. (1996). *Basic Concepts of Legal Thought*. Oxford: Oxford University Press.

Foot, P. (1967/2002). The problem of abortion and the doctrine of the double effect. In P. Foot, *Virtues and Vices and Other Essays in Moral Philosophy* (pp. 19-32). Oxford: Oxford University Press.

Freitas, J. D., Anthony, S. E., & Alvarez, G. (2019). Doubting driverless dilemmas. Retrieved January 26, 2019, from https://psyarxiv.com/a36e5/

Fried, C. (1970). *Anatomy of Value*. Cambridge, MA: Harvard University Press.

Friedman, L. M., & Arold., N.-L. (2011). Cannibal rights: A note on the modern law of privacy. *Northwestern Interdisciplinary Law Review*, 235.

Frowe, H. (2010). A practical account of self-defence. *Law and Philosophy*, *29*(3), 245-272.

Frowe, H. (2014). *Defensive Killing*. Oxford: Oxford University Press.

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Oxford University Press.

Geistfeld, M. A. (2017). A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation. *California Law Review*(105), 1611.

Gerdes, J. C., & Thornton, S. M. (2015). Implementable ethics for autonomous vehicles. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren* (pp. 87-102). Springer.

Gibbard, A. (1990/2002). *Wise Choices, Apt Feelings*. New York: Oxford University Press.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, 681-700.

Goodall, N. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 58-65.

Goodall, N. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, *30*(8), 810-821.

Goodall, N. (2017). From trolleys to risk: Models for ethical autonomous driving. *American Journal of Public Health*, *107*(4), 496.

Goodall, N. (2019). More than trolleys: Plausible, ethically ambiguous scenarios likely to be encountered by automated vehicles. *Transfers: Interdisciplinary Journal of Mobility Studies*, *9*(2), 45-58.

Graham, P. A. (2010). In defense of objectivism about moral obligation. *Ethics*, *121*(1), 88-115.

Greene, J. D. (2008). The secret joke of Kant's soul. *Moral Psychology*, *3*, 35-79.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

Gurney, J. K. (2016). Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law. *Albany Law Review*, *79*(1), 183-267.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist*, *102*(2), 259-275.

Hammond, P. (1976). Equity, Arrow's conditions, and Rawls' difference principle. *Econometrica: Journal of the Econometric Society*, *44*(4), 793-804.

Hannawald, L. a. (2004). Equal effectiveness study on pedestrian protection. *Technische Universität Dresden*.

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology*, *11*(1), 91-99.

Haque, A. A. (2012). Killing in the fog of war. *Southern California Law Review*, *86*, 63-116.

Hare, R. M. (2002). *Essays on Bioethics*. Oxford: Oxford University Press.

Harsanyi, J. (1977). Rule utilitarianism and decision theory. *Erkenntnis*, *11*, 25-53.

Harsanyi, J. (1982). Morality and the theory of rational behaviour. In *Utilitarianism and Beyond* (pp. 32-62). Cambridge: Cambridge University Press.

Harsanyi, J. (1993). Expectation effects, individual utilities, and rational desires. In B. Hooker (Ed.), *Rationality, Rules, and Utility: New Essays on the Moral Philosophy of Richard Brandt* (pp. 115-126). Boulder, CO: Westview Press.

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 434-435.

Hart, H. (1961). *The Concept of Law*. Oxford: Clarendon.

Hayenhjelm, M., & Wolff, J. (2012). The moral problem of risk impositions: A survey of the literature. *European Journal of Philosophy*, *20*, E26-E51.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135-175.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley.

Hevelke, A. &.-R. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics*, *21*, 619-630.

Hicks, J. (1939). The foundations of welfare economics. *The Economic Journal*, *49*(196), 696-712.

Hill v Baxter [1958] 1 QB 277.

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, *21*(3), 669-684.

Hirose, I. (2015). *Moral Aggregation*. New York: Oxford University Press.

Hobbes, T. (1651/2008). *Leviathan*. Oxford: Oxford University Press.

Hohfeld, W. N. (1919). *Fundamental Legal Conceptions*. (W. W. Cook, Ed.) New Haven: Yale University Press.

Holm, S. A. (2016). Right against risk-imposition and the problem of paralysis. *Ethical Theory and Moral Practice*, *19*, 917-930.

Hooker, B. (1990). Rule consequentialism. *Mind*, *99*(393), 67-77.

Hooker, B. (1996). Ross-style pluralism versus rule-consequentialism. *Mind*, *105*(420), 531-552.

Hooker, B. (2000). *Ideal Code, Real World*. Oxford: Oxford University Press.

Hooker, B. (2003). Moral particularism: bad and wrong. In B. Hooker, & M. O. Little (Eds.), *Moral Particularism* (pp. 1-22). Oxford: Oxford University Press.

Horty, J. (2007). Reasons as Defaults. *Philosophers' Imprint*, *7*.

Hübner, D. &. (2018). Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice*, *21*(3), 685-698.

Hurka, T. (2016). Trolleys and permissible harm. In F. M. Kamm, *The Trolley Problem Mysteries: The Berkely Tanner Lectures* (pp. 135-50). New York, NY: Oxford University Press.

Husak, D. N. (1994). Is drunk driving a serious offense? *Philosophy and Public Affairs*, *23*(1), 52-73.

Jackson, F. (1991). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*, *101*(3), 461-82.

Jackson, F., & Smith, M. (2006). Absolutist moral theories and uncertainty. *The Journal of Philosophy*, *3*(6), 267-283.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 195-204.

Joyce, R. (2008). What neuroscience can (and cannot) contribute to metaethics. *Moral Psychology, 3*, 371-394.

Kagan, S. (1988). The additive fallacy. *Ethics, 99*(1), 5-31.

Kagan, S. (1989). *The Limits of Morality*. Oxford: Oxford University Press.

Kagan, S. (1992). The structure of normative ethics. *Philosophical Perspectives, 6*, 223-242.

Kagan, S. (1998). *Normative Ethics*. Oxford: Westview Press.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu., J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, 193-209.

Kaldor, N. (1939). Welfare propositions in economics and interpersonal comparisons of utility. *The Economic Journal, 49*(195), 549-552.

Kamm, F. (1993). *Morality, Mortality, Volume I: Death and Whom to Save From It*. New York: Oxford University Press.

Kamm, F. M. (2007). *Intricate Ethics*. Oxford: Oxford University Press.

Kamm, F. M. (2016). *The Trolley Problem Mysteries*. New York: Oxford University Press.

Kant, I. (1785/2002). *Groundwork for the Metaphysics of Morals*. Binghamton, NY: Vail-Ballou Press.

Kauppinen, A. (Forthcoming). Who should bear the risk when self-driving vehicles crash? *Journal of Applied Philosophy*.

Keeling, G. (2017). Commentary: Using virtual reality to assess ethical decisions in road traffic scenarios: applicability of value-of-life-based models and influences of time pressure. *Frontiers in Behavioral Neuroscience, 11*, 247-8.

Keeling, G. (2018a). Legal necessity, Pareto efficiency & justified killing in autonomous vehicle collisions. *Ethical Theory and Moral Practice, 21*(2), 413-427.

Keeling, G. (2018b). Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2017*. Berlin: Springer.

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics, 26*, 293-307.

Keeling, G., Evans, K., Thornton, S. M., Mecacci, G., & Santoni de Sio, F. (2019). Four Perspectives on What Matters for the Ethics of Automated Vehicles. In G. Meyer, & S. Beiker (Eds.), *Road Vehicle Automation 6*. Springer International Publishing.

Kitcher, P. (1993). *The Advancement of Science: Science without Legend, Objectivity without Illusions*. New York: Oxford University Press.

Kramer, M. (2009). *Moral Realism as a Moral Doctrine*. Chichester: John Wiley and Sons.

Krantz, D., Duncan Luce, R., Suppes, P., & Tversky, A. (1971). *The Foundations of Measurement. Volume 1: Additive and Polynomial Representations*. New York: Academic Press.

Lang, G. (2007). The limits of self-defence (Unpublished Manuscript).

Lazar, S. (2009). Responsibility, risk, and killing in self-defense. *Ethics, 119*(4), 699-728.

Lazar, S. (2012). Necessity and self-defence in war. *Philosophy and Public Affairs, 40*(1), 3-44.

Lazar, S. (2018). In dubious battle: uncertainty and the ethics of killing. *Philosophical Studies, 175*, 859-883.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19*(2), 107-115.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell Publishers.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Nous, 13*(4), 455-476.

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving: Technical, Legal and Social Aspects* (pp. 69-85). Springer Berlin Heidelberg.

Litman, T. (2020). *Autonomous Vehicle Implementation Predictions: Implications for Transport Planning*. Victoria Transport Policy Institute.

Lockhart, T. (2000). *Moral Uncertainty and its Consequences*. Oxford: Oxford University Press.

Lord, E., & Maguire, B. (2016). An opinionated guide to the weight of reasons. In E. Lord, & B. Maguire (Eds.), *Weighing Reasons* (pp. 3-24). Oxford: Oxford University Press.

MacAskill, W. (2016). Normative uncertainty as a voting problem. *Mind*, *125*(500), 967-1004.

MacAskill, W., & Ord., T. (2018). Why maximize expected choice-worthiness? *Noûs*, 1-27.

MacKay, D. J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

Mackie, G. (2003). *Democracy Defended*. Cambridge: Cambridge University Press.

Manson, N. C., & O'Neill, O. (2007). *Rethinking Informed Consent in Bioethics*. New York: Cambridge University Press.

Marchant, G. E., & Lindor, R. A. (2012). The coming collision between autonomous vehicles and the liability system. *Santa Clara Law Review*, *52*, 1321.

Mason, E. (2012). Objectivism and prospectivism about rightness. *Journal of Ethics and Social Philosophy*, *7*(2).

Mawhinney, G. R. (2013). To be ill or to kill: The criminality of contagion. *The Journal of Criminal Law*, *77*(3), 202-214.

McCarthy, D. (1997). Rights, explanation, and risks. *Ethics*, *107*, 205-225.

McMahan, J. (1994). Innocence, self-defence and killing in war. *Journal of Political Philosophy*, *2*(3), 193-221.

McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press.

McMahan, J. (2005). The basis of moral liability to defensive killing. *Philosophical Issues*, *15*, 386-405.

McMahan, J. (2009). *Killing in War*. Oxford: Oxford University Press.

McMahan, J. (2011). Duty, obedience, desert, and proportionality in war: A response. *Ethics*, *122*, 135-167.

McNaughton, D. (1996/2013). An unconnected heap of duties? In R. Shafer-Landau, *Ethical Theory: An Anthology* (pp. 763-771). Chichester: John Wiley and Sons.

Millar, J. (2014). Technology as moral proxy: Autonomy and paternalism by design. *IEEE Ethics in Engineering, Science and Technology Proceedings.*

Minsky, M., & Papert, S. (1969). *Perceptrons.* Cambridge, MA: MIT Press.

Mouse's Case (1608) Michaelmas Term 6, JMS I. vol 12.

Nagel, T. (1979). Fragmentation of Value. In T. Nagel, *Mortal Questions* (pp. 128-41). Cambridge: Cambridge University Press.

Nagel, T. (1986). *The View from Nowhere.* Oxford University Press: New York.

Nair, S. (2016). How Do Reasons Accrue? In E. Lord, & B. Maguire (Eds.), *Weighing Reasons* (pp. 56-73). Oxford: Oxford University Press.

Narverson, J. (1973). Moral problems of population. *The Monist, 57*(1), 62-86.

Narveson, J. (1988). *The Libertarian Idea.* Philadelphia, PA: Temple University Press.

National Highway Traffic Safety Administration (NHTSA) Center for Statistics and Analysis. (2016). *Traffic safety facts: 2014 data.* Retrieved April 13, 2020, from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812262

National Transportation Safety Board (NTSB). (2018). *Preliminary Report Highway HWY18MH010.*

Negishi, T. (1960). Welfare economics and the existence of an equilibrium for a competitive economy. *Metroeconomica, 12*(2-3), 92-97.

Nezami, F. N., Wächter, M. A., Pipa, G., & König, P. (2020). Project Westdrive: Unity city with self-driving cars and pedestrians for virtual reality studies. *Frontiers in ICT, 7*, 1.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625-632.

Norcross, A. (1997). Comparing harms: headaches and human lives. *Philosophy and Public Affairs, 26*(2), 135-167.

Norcross, A. (2008). Off her trolley? Frances Kamm and the metaphysics of morality. *Utilitas*, *20*(1), 65-80.

Nozick, R. (1974). *Anarchy, State and Utopia*. New York: Basic Books.

Nozick, R. (2001). *Invariances*. Cambridge, MA: Harvard University Press.

Nyholm, S. (2018a). The ethics of crashes with self-driving cars: a roadmap I. *Philosophy Compass*, 1-10.

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, *13*(7), e12506.

Nyholm, S. (2018c). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *4*, 1201-1219.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, *19*(5), 1275-1289.

Oberdiek, J. (2009). Towards a right against risking. *Law and Philosophy*, *28*(4), 367-392.

Oberdiek, J. (2012). The moral significance of risking. *Legal Theory*, *18*, 339-356.

Oddie, G. (1994). Moral uncertainty and human embryo experimentation. In K. W. Fulford, G. Gillett, & J. M. Soskice (Eds.), *Medicine and Moral Reasoning* (pp. 144-161). Cambridge: Cambridge University Press.

Offences Against the Person Act. (1861).

Oh, C. K. (2008). Development of probabilistic pedestrian fatality model for characterizing pedestrian-vehicle collisions. *International Journal of Automotive Technology*, *9*(2), 191-196.

Okasha, S. (2011). Theory choice and social choice: Kuhn versus Arrow. *Mind*, *120*(447), 83-115.

O'Neill, O. (2003). Autonomy: the emperor's new clothes. *Aristotelian Society Supplementary Volume*, *77*(1), 1-21.

Ori, M. (2014). The morality of motorcycling. *Philosophical Papers*, *43*(3), 345-363.

Ori, M. (2015). Motorcycling as a moral improvement: A response to Hansson. *Philosophical Papers*, *44*(3), 377-387.

Otsuka, M. (1994). Killing the innocent in self-defense. *Philosophy and Public Affairs*, *23*(1), 74-94.

Otsuka, M. (2006). Saving lives, moral theory, and the claims of individuals. *Philosophy and Public Affairs*, *34*(2), 109-135.

Otsuka, M. (2012). Prioritarianism and the separateness of persons. *Utilitas*, *24*(3), 365-380.

Otsuka, M., & Voorhoeve, A. (2009). Why it matters that some are worse off than others: an argument against the priority view. *Philosophy and Public Affairs*, *37*(2), 171-199.

Parfit, D. (1987). *Reasons and Persons*. Oxford: Oxford University Press.

Parfit, D. (2003). Justifiability to each person. *Ratio*, *16*(4), 368-390.

Parfit, D. (2011a). *On What Matters* (Vol. 1). Oxford: Oxford University Press.

Parfit, D. (2011b). *On What Matters* (Vol. 2). Oxford: Oxford University Press.

Pasanen, E. (1992). *Driving speeds and pedestrian safety: A Mathematical Model*. Helsinki: Helsinki University of Technology, Finland.

Plato. (2002). Euthyphro. In *Five Dialogues* (G. M. Grube, Trans., 2nd ed., pp. 1-20). Indianapolis, IN: Hackett.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, *10*(3), 61-74.

Porta, J. M., Vlassis, N., Spaan, M. T., & Poupart, a. P. (2006). Point-based value iteration for continuous POMDPs. *Journal of Machine Learning Research*, 2329-2367.

Prakken, H. (2005). A Study of Accrual of Arguments. *Proceedings of Tenth International Conference on Artificial Intelligence and Law* (pp. 85-94). New York: ACM Press.

Prichard, H. A. (2002/1932). Duty and Ignorance of Fact. In *Moral Writings* (pp. 84-101). Oxford: Clarendon Press.

Pritchard, D. (2015). Risk. *Metaphilosophy*, *46*(3), 436-461.

Pritchard, D. (2016). Epistemic risk. *Journal of Philosophy*, *113*(11), 550-571.

Quong, J., & Firth, J. (2012). Necessity, moral liability, and defensive harm. *Law and Philosophy, 31*, 673-701.

R v Bateman [1925] 19 Cr App R 8.

R v Brown [1994] 1 AC 212.

R v Caldwell [1982] AC 341.

R v Cunningham [1957] 2 QB 396.

R v Dica [2004] EWCA Crim 1103.

R v Dudley and Stephens [1884] QBD 273.

R v Emmett [1999] All ER (D) 641 (CA).

R v G [2003] UKHL 50.

R v G [2008] UKHL 37.

R v Konzani [2005] EWCA Crim 706.

R v Lawrence 1982 AC 510.

R v Martin [2001] 1 Cr App R 27.

R v Matthews and Alleyne [2003] Cr App R 30.

R v Wilson [1996] 3 WLR 125.

R v Woolin [1999] 1 A.C. 82.

Random Science Tools. (2017, December 04). Car Stopping Distance Calculator. Retrieved April 20th, 2020, from https://www.random-science-tools.com/physics/stopping-distance.htm

Rasmussen, K. (2012). Should the probabilities count? *Philosophical Studies, 159*(2), 205-218.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.

Rawls, J. (2005). *Political Liberalism: Expanded Edition*. New York: Columbia University Press.

Raz, J. (2003). The truth in particularism. In B. Hooker, & M. O. Little (Eds.), *Moral Particularism* (pp. 48-78). Oxford: Oxford University Press.

Re A (Conjoined twins) [2001] 2 WLR 480 (CA).

Reninger v Fagossa [1551] 1 Plowd. 1, 75 Eng. Rep. 1..

Rivera-López, E. (2008). Probabilities in tragic choices. *Utilitas*, *20*(3), 323-333.

Road Traffic Act. (1930).

Rodin, D. (2002). *War and Self-Defense*. Oxford: Clarendon Press.

Rodin, D. (2011). Justifying harm. *Ethics*, *122*(1), 74-110.

Rosén, E. a. (2009). Pedestrian fatality risk as a function of car impact speed. *Accident Analysis & Prevention*, *41*(3), 536-462.

Ross, J. (2006). Rejecting ethical deflationism. *Ethics*, 742-68.

Ross, W. D. (1930/2007). *The Right and the Good*. (P. Stratton-Lake, Ed.) New York: Oxford University Press.

Ross, W. D. (1939). *The Foundations of Ethics*. Oxford: Oxford University Press.

Ruscio, D., Ciceri, M. R., & Biassoni., F. (2015). How does a collision warning system shape driver's brake response time? The influence of expectancy and automation complacency on real-life emergency braking. *Accident Analysis & Prevention*, *77*, 72-81.

Salmon, W. C. (1999). Scientific Explanation. In M. H. Salmon, J. Earman, C. Glymour, J. G. Lennox, P. Machamer, J. E. McGure, . . . K. F. Schaffner, *Introduction to the Philosophy of Science* (pp. 7-41). Indianapolis, IN: Hackett.

Santoni de Sio, F. (2017). Killing by Autonomous Vehicles and the Legal Doctrine of Necessity. *Ethical Theory and Moral Practice*, *20*(2), 411-429.

Savage, L. (1972). *The Foundations of Statistics*. New York: Dover.

Scanlon, T. (1998). *What we Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.

Scanlon, T. (2011). How I am not a Kantian. In D. Parfit, *On What Matters Vol. 2* (pp. 116-139). Oxford: Oxford University Press.

Scheffler, S. (1988). Introduction. In S. Scheffler (Ed.), *Consequentialism and Its Critics* (pp. 1-13). New York: Oxford University Press.

Schelling, T. (2006). Should the numbers determine whom to save? In *The Strategies of Commitment* (pp. 113-146). Cambridge, MA: Harvard University Press.

Schroeder, C. H. (1986). Rights against risks. *Columbia Law Review*, 495-562.

Sen, A. (1970). *Collective Choice and Social Welfare*. San Francisco: Holden Day.

Sen, A. (1976). Welfare inequalities and Rawlsian axiomatics. *Theory and Decision*, *7*(4), 243-262.

Sen, A. (2011). *The Idea of Justice*. London: Penguin Books.

Sepielli, A. (2009). What to do when you don't know what to do. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 4, pp. 5-28). Oxford: Oxford University Press.

Sexual Offences Act. (2003).

Shafer-Landau, R. (1994). Ethical disagreement, ethical objectivism and moral indeterminacy. *Philosophy and Phenomenological Research*, *54*(2), 331-344.

Sidgwick, H. (1907/1981). *The Methods of Ethics*. Indianapolis, IN: Hacket Publishing Company.

Smith, A. (1759/2009). *The Theory of Moral Sentiments*. New York, NY: Penguin Books.

Smith, B. W. (2012, March). Driving at Perfection. The Center for Internet and Society at Stanford Law School. Retrieved April 13th, 2020, from http://cyberlaw.stanford.edu/blog/2012/03/driving-perfection

Smith, B. W. (2017). Automated driving and product liability. *Michigan State Law Review*, 1.

Smith, T., & Simmons, a. R. (2012). Heuristic search value iteration for POMDPs. Retrieved from arXiv preprint arXiv:1207.4166

Society for Automotive Engineers (SAE). (2014). *Information Report J3016: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems.*

Steinbock, B. (1985). Drunk driving. *Philosophy and Public Affairs*, *14*(3), 278-295.

Sucar, L. E. (2015). *Probabilistic Graphical Models: Principles and Applications*. London: Springer.

Sullins, J. P. (2006). When is a robot a moral agent? *Machine Ethics*, 151-60.

Sweet v Parsley [1970] AC 132.

Tadros, V. (2011). *The Ends of Harm: The Moral Foundations of the Criminal Law*. Oxford: Oxford University Press.

Tadros, V. (2016). *Wrongs and Crimes*. Oxford: Oxford University Press.

Taurek, J. M. (1977). Should the numbers count? *Philosophy and Public Affairs*, 293-316.

Thomson, J. J. (1976). Killing, letting die and the trolley problem. *The Monist, 59*, 204-217.

Thomson, J. J. (1990). *The Realm of Rights*. Cambridge, MA: Harvard University Press.

Thomson, J. J. (1991). Self-defense. *Philosophy and Public Affairs, 20*(4), 283-310.

Thomson, J. J. (2008). Turning the trolley. *Philosophy and Public Affairs, 36*(4), 359-374.

Thomson, J. J. (2016). Kamm on Trolley Problems. In F. M. Kamm, *The Trolley Problem Mysteries* (pp. 113-134). Oxford: Oxford University Press.

Uniacke, S. (1994). *Permissible Killing: The Self-Defence Justification of Homicide*. Cambridge: Cambridge University Press.

Urmson, J. (1975). A defence of intuitionism. *Proceedings of the Aristotelian Society*, 111-119.

von Neumann, J., & Morgenstern, O. (1990). *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.

Voorhoeve, A., & Fleurbaey, M. (2012). Egalitarianism and the separateness of persons. *Utilitas, 24*(3), 381-298.

Walzer, M. (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. New York: Basic Books.

Weatherson, B. (2014). Running risks morally. *Philosophical Studies, 167*(1), 141-163.

Wood, A. (1999). *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *International Conference on Machine Learning, 1*, 609-616.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694-699.

Zimmerman, M. J. (2008). *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge: Cambridge University Press.

Zohar, N. J. (1993). Collective war and individualistic ethics: Against the conscription of "self-defense". *Political Theory*, *21*(4), 606-622.